Lessons from SSA Demonstrations for Disability Policy and Future Research

Edited by

Austin Nichols ■ Jeffrey Hemmeter ■ Debra Goetz Engler

SOCIAL SECURITY ADMINISTRATION
USA

Abt ASSOCIATES

Lessons from SSA Demonstrations for Disability Policy and Future Research

Lessons from SSA Demonstrations for Disability Policy and Future Research

Austin Nichols
Jeffrey Hemmeter
Debra Goetz Engler
Editors

2021

ISBN: 979-8-9851403-1-6

Sponsored by the Social Security Administration through a contract with Abt Associates. (Contract #28321320C00060010)

For more information about the meeting at which the content of this volume was first presented, see https://ssa-demonstration-lessons.abtassociates.com/.

The views expressed in this volume are those of the authors and do not necessarily represent the views of the Social Security Administration or the US federal government. The authors take responsibility for all statements in their respective chapters.

Published by Abt Press
6130 Executive Blvd.
Rockville, MD 20852

Suggested citation: Nichols, Austin, Jeffrey Hemmeter, and Debra Goetz Engler, eds. 2021. *Lessons from SSA Demonstrations for Disability Policy and Future Research.* Rockville, MD: Abt Press.

Contents

Exhibits

Acknowledgments

The editors of this volume would like to acknowledge the efforts of many individuals who contributed to the *Lessons Learned from SSA Demonstrations: A State of the Science Meeting* convened on June 15, 2021, for which these papers were prepared and presented. We appreciate the participation of more than 400 disability and social policy experts, researchers, program operators, and advocates who attended the virtual event.

We are grateful for the insights generated at the meeting about the lessons from the Social Security Administration's past demonstrations and directions for the future. We also wish to acknowledge the many individuals who have contributed to this volume. They endured several rounds of comments and revisions to produce this unified volume, while still retaining their individual voices.

At SSA, we want to thank Joyanne Cobb, Jackson Costa, Kai Filion, Nitin Jagdish, John Jones, Marion McCoy, Joyce Nicholas, Paul O'Leary, Alexander Strand, Terri Uttermohlen, Robert Weathers, and Susan Wilschke for their careful review and insightful comments on earlier drafts of the volume chapters. Additionally, we would like to thank Mark Warshawsky, Kilolo Kijakazi, Katherine Bent, and Susan Wilschke for their support of the project.

At Abt Associates, we wish to acknowledge the work of Laura Peck, who provided technical review and comments on the volume chapters and editorial input and guidance for the layout of the volume. We are also grateful for the contributions of Daniel Gubits and Sarah Prenovitz, who worked with chapter authors to review initial outlines and drafts of the chapters. Michelle Wood provided management and technical oversight to the Abt team. We are grateful to Bry Pollack for copy editing the volume; to Brittany Tuwamo for collating and checking the references and other editorial assistance; and to Erin Miles for coordinating production and formatting the volume. The Hatcher Group designed the book cover. Jennifer Bagnell Stuart managed the project with assistance from Cara Sierks.

Finally, we want to thank the dedicated and industrious subject matter experts at SSA who have administered its demonstrations over the years, creating and overseeing miniature government programs with limited resources.

Foreword

Kilolo Kijakazi
Acting Commissioner, Social Security Administration

The Social Security Administration (SSA) administers two of the most effective antipoverty programs for people with disabilities—the Social Security Disability Insurance (SSDI) program and the Supplemental Security Income (SSI) program. Each year, we serve millions of individuals with disabilities in these two programs, providing the cash supports needed to help them be economically secure. Although these programs are effective, it is important to continually explore policy changes to help them become better. The demonstrations reviewed in *Lessons from SSA Demonstrations for Disability Policy and Future Research* cover a wide array of policy tests. The review provided in this volume will help us determine which policies we already have sufficient information about and which policies need additional evidence. It will also help identify ways to better implement these types of projects so that they provide useful information to policymakers.

One of the most important questions we need more evidence about is whether the programs are equitable. As the papers in this volume note, we know a lot about the aggregate effects of policies. But that is insufficient for truly understanding whether the programs are effectively addressing barriers—barriers to employment, health, well-being, and other social goals. These questions should not just be about whether people of color or people with specific impairments have smaller or larger impacts. That is important, but we need to know more. We need to know *why* these impacts are smaller or larger. These questions need to be included in the design of the demonstrations. President Biden's Executive Order, on advancing racial equity, mandates that we consider these types of questions. We need to think more about structural barriers in the labor market and service sectors. My own research, and that of others, has pointed to structural racism as a prominent barrier for people of color.

One issue that needs to be addressed is the quality of data available to answer these questions. SSA is in the process of revisiting how we collect and use administrative data on race and ethnicity. Having this information will be important to answering these questions. For too long SSA and other agencies have relied on the thought that the only data that can be defined as programmatic—and thus we have the need or ability to collect—are the data that go into decision making and program rules. We need to decide that ensuring equitable outcomes is part of the programmatic needs of government systems—that it is part of what these programs are intended to accomplish. There are too many examples of policies and programs that have been shown to be inequitable.

Asking additional questions will only get us so far. The teams that worked on these demonstrations include high-quality economists, statisticians, program evaluators, and social scientists. However, we also need to make sure that teams are

inclusive of the people with lived experiences in the issues studied so that the questions being asked and assumptions about what will work have a basis in reality.

None of the observations mentioned above take away from the work that has been done to produce rigorous evidence through SSA's demonstrations, but they do mean that we can and should take further action to improve our SSI and SSDI programs. Effective research must examine the systemic barriers that impact individuals and engage with individuals to explore solutions. This volume is a positive step in that direction, and future demonstrations can help make further advances.

Chapter 1

An Introduction to Disability Policy and SSA's Demonstrations

Austin Nichols
Abt Associates
Jeffrey Hemmeter
Social Security Administration[1]

The Social Security Administration (SSA) administers the two largest federal disability insurance programs in the United States: Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). Within the scope of its congressionally mandated authority, SSA makes a variety of policy choices that affect the economic well-being of SSDI beneficiaries and SSI recipients. This includes promoting the employment of beneficiaries and recipients through a multitude of work incentives policies and programs.

Over the past four decades, SSA has also conducted many tests of new policies and programs to improve participants' outcomes. These tests, called "demonstrations," address many policy-relevant topics including family supports, health insurance, transition to adulthood, informational notices, changes to benefit calculations, and a variety of employment services and waivers of program rules, as detailed in the Appendix.

At the outset, it is useful to distinguish a demonstration from an intervention or evaluation. An *intervention* is a policy or program change intended to affect participant outcomes; an intervention may or may not be evaluated. A *demonstration* is a temporary intervention or a package of interventions of limited scale (i.e., not rolled out nationwide or to all beneficiaries or recipients), implemented for the purpose of being evaluated. An *evaluation* generates the information by which a demonstration can inform decisions about whether the tested intervention (or some version of it) should be implemented permanently or more broadly.

SSA's demonstrations have generated dozens of documents reporting how policies and programs worked, and for whom. However, in 2004, the Government Accountability Office (GAO 2004) critiqued the impact of the demonstrations, stating that "SSA's demonstration projects have had little impact on the agency's and the Congress' consideration of [disability insurance] policy issues" (3). GAO reported that even though "SSA has used methodological designs that GAO determined were strong or reasonable when assessed against professional research standards for 11 of its 14 projects,…these projects have yielded limited information on the impacts of the

[1] The views expressed in this chapter are those of the authors and do not necessarily represent the views of the Social Security Administration or the US federal government.

program and policy changes they were testing" (2008, inside cover). As of its 2008 report, of the 14 projects GAO reviewed, 5 had been completed and 5 canceled. SSA subsequently instituted new policies to improve future demonstrations, including those covered in this volume.

This volume synthesizes the findings of many of SSA's demonstrations to identify cross-demonstration lessons about which policies, programs, and other operational decisions could provide effective supports for SSDI beneficiaries and SSI recipients who want to work. It also identifies lessons for the design and use of demonstrations for future learning. This chapter provides an overview of the SSDI and SSI programs, SSA's demonstration portfolio, and selected lessons from the remaining chapters in the volume.

By taking stock of the lessons learned from prior demonstrations, policymakers can better understand what has been tested and whether and why the tested interventions were effective. These demonstrations have informed policy discussions and proposals, although not always in expected ways, and this volume brings together findings that typically have been discussed in isolation. This synthesis will enable SSA and other stakeholders to implement policies and programs that work in multiple settings, propose alternatives to them that might not have worked for identifiable reasons, and identify policies and strategies to test in future demonstrations.

FEDERAL DISABILITY INSURANCE PROGRAMS

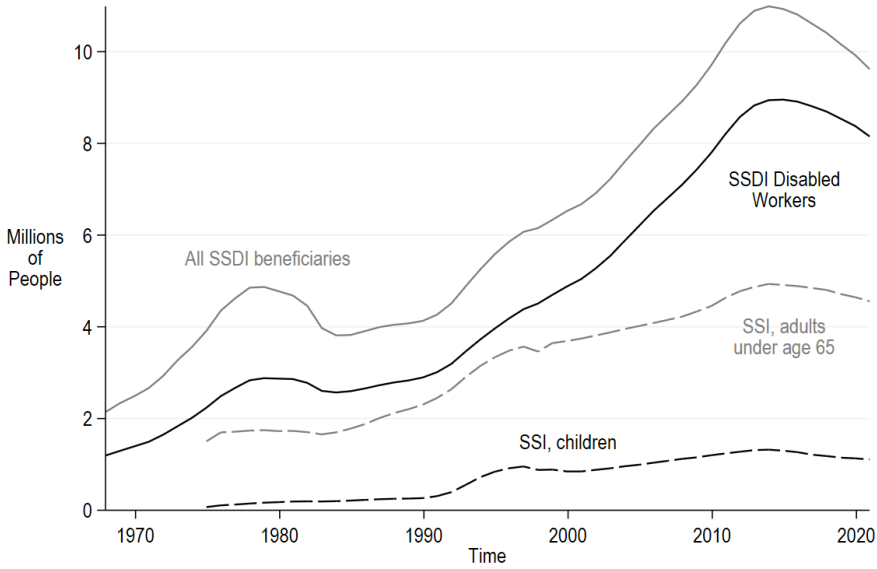
The disability programs run by SSA paid more than $130 billion to 9.6 million SSDI beneficiaries in 2020, $9.6 billion to 1.1 million youth receiving SSI, and $35.7 billion to 4.5 million adult SSI recipients younger than age 65. Most SSI recipients have no other income sources (CBPP 2021), and half of SSDI beneficiaries and about two-thirds of SSI recipients have income less than the poverty threshold when SSDI benefits and SSI payments are not included (Bailey and Hemmeter 2015). These programs provide economic security to people with barriers, enabling them to pay for food, housing, and other necessities if they are unable to work or rely on somebody who is unable to work due to a disability. Further, SSDI benefit receipt is associated with reduced mortality (Gelber, Moore, and Strand 2017) and SSI receipt is associated with improved childhood outcomes (Guldi et al. 2018).

In December 2020, nearly 10 million individuals, of whom 8.15 million were disabled worker beneficiaries, received monthly SSDI payments averaging $1,260.[2] Eight million individuals received monthly SSI payments averaging $576, of whom

[2] SSDI recipients can be categorized as Disabled Workers, who receive benefits based on their own earnings, or as Disabled Adult Children or Disabled Widow(er)s, who receive benefits based on a parent's or spouse's earnings.

4.56 million were adults younger than age 65 and 1.11 million were children.[3] About 11 percent of SSDI beneficiaries also received SSI. The numbers of total SSDI beneficiaries, the number who are disabled workers, and SSI recipients younger than age 65, both adults and children, in December of each year since 1967, are shown in Exhibit 1.1.

Exhibit 1.1. All SSDI Beneficiaries and "Disabled Workers," SSI Recipients Adults Younger than Age 65, and Children, 1967-2020



Source: Data maintained by SSA, reported in *Monthly Statistical Snapshots* and annual statistical supplements to the *Social Security Bulletin*.
Note: SSI was established in 1972, with payments beginning in 1974. Numbers reported here are only for federally administered payments, including some state-supplement-only recipients, but not SSI recipients receiving no federal payments.

Liebman (2015) characterized changes in SSDI participation as due primarily to increased eligibility among women and declining mortality among SSDI beneficiaries. Between 1993 and 2007, he finds that population aging and increased eligibility among women accounted for two-thirds of the increase in SSDI benefit receipt, with rising incidence among women accounting for one-quarter and declining mortality accounting for one-sixth. He concludes, "the case for [disability insurance] reform is not primarily a fiscal one—up until the 2007–2009 recession, spending on the

[3] Payments for child recipients averaged $675 and payments to adults younger than age 65 averaged $606, whereas payments to older adults averaged $468, bringing down the overall average. For more detail, see the *SSA Monthly Statistical Snapshots* and annual statistical supplements to the *Social Security Bulletin*.

program…had increased by only 0.13 percent of [Gross Domestic Product] over 30 years" (124).

Policymakers have had an interest in the growth in the SSDI and SSI programs in the last several decades, and particularly in the relationship between disability and work. Scholars (e.g., Parsons 1980) have long argued that the availability of disability benefits lowers labor force participation among program participants. Bound and colleagues (1989, 1991, 2003, 2014), among others, provide evidence suggesting that the availability of disability benefits or changes in policy do not fully explain declines in labor force participation. Parallel to a focus on work outcomes, demonstrations have examined impacts on health and other dimensions of well-being. This led to legislation and policies (and demonstrations) focused on supporting beneficiaries' efforts to return to work.

The Social Security Disability Amendments of 1980 established the Extended Period of Eligibility for SSDI and added several other work incentives to the SSDI and SSI programs. The Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act) created new programs for individuals who receive disability benefits. These and other policy changes were motivated by the idea that many SSI recipients and SSDI beneficiaries have the capacity and desire to work. However, use of these work incentives has been low. Because of widespread concern that the availability of disability benefits could lead some beneficiaries not to work, and the perception that many beneficiaries might prefer to work given the incentive to do so, most SSA demonstrations have tested how to encourage work among persons with disabilities. Whether the disability benefits crowding out work or the strict definition of disability for SSDI and SSI is the reason for low return-to-work outcomes is an ongoing debate.

The Disability Insurance Trust Fund has repeatedly been in imminent danger of depletion, most recently in 2016 when its depletion was averted by a temporary repurposing of funds from the Old-Age and Survivors Insurance Trust Fund and a downturn in the number of beneficiaries. The long-term consequences of the pandemic and recession that began in 2020 could present new challenges to the programs. Currently, the Disability Insurance Trust Fund is expected to be depleted in 2057 (Board of Trustees 2021).

The growth in children receiving SSI benefits in the 1990s, particularly for mental impairments, also spurred interest in that program. However, recent research has suggested that the growth of SSI may have been less than would have been expected given the growth in the poverty rate and number of children with disabilities (NASEM 2015).

Work Disability Insurance, Not Disability Insurance

Even though SSDI is explicitly an insurance program and SSI is a transfer program that is not technically insurance, both are forms of the broad concept of social insurance that protects individuals against the loss of earning capacity. It is important to note that these SSA programs insure work disability, not any kind of disability. That

is, a functional impairment that does not affect an adult's ability to work will not result in an award of benefits. Like other forms of insurance, the optimal amount of insurance balances marginal well-being across people's different potential experiences, or states of the world, but it is not intended to make up for any loss that could be incurred.

The Social Security Act[4] Sec. 223(d)(1)(A) defines this work disability for adults:

> inability to engage in substantial gainful activity [SGA] by reason of any medically determinable physical or mental impairment which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months.

In 2020, SSA considered an individual earning at least $1,260 a month (or $2,110, for statutorily blind individuals) as engaging in SGA. This is adjusted for inflation annually, and in 2021 the SGA threshold is $1,310 a month (or $2,190, for statutorily blind individuals).

Note that this standard involves an individual's functional capacity, their job qualifications such as work history and completed education, and attributes of the labor market. A change in the labor market that led to someone being unable to engage in SGA, such as recessions or technological displacement, even with no change in functional limitations, could mean that person now has a qualifying disability. As a result, one should expect work disability rates to rise whenever wages decline or job opportunities dry up (Autor and Duggan 2000; Black, Daniel, and Sanders 2002; Charles, Li, and Stephens 2018; Nichols, Schmidt, and Sevak 2017; Vachon 2014). This is sometimes framed as induced entry by economic conditions, but is inherent in the statutory definition of disability.

Like other forms of social insurance in the United States, SSDI and SSI are not designed to make individuals with disabilities as well off as they were before the onset of disability, but designed to mitigate the personal and societal losses in an equitable manner. Numerous authors have sought to ascertain whether SSDI and SSI policies are designed to attain the greatest net benefit to society or could be improved (e.g., Bound et al. 2004).

The costs of providing social insurance arise from the full social cost of providing cash and noncash benefits, which includes distortions to labor markets (Chetty 2006). The benefits include the value of insurance (Eeckhoudt and Kimball 1992; Kimball 1990) and the value of redistribution (Finkelstein and Hendren 2020; Hendren 2016, 2020; Hendren and Sprung-Keyser 2019). These economic ideas about social insurance and its value are quite distinct from the accounting framework typically adopted in a benefit-cost calculation of an intervention, but rough adjustments are often made to account for factors such as distributional effects or opportunity costs. For example, in the Benefit Offset National Demonstration (BOND), the benefit-cost

[4] There is a special definition for people who are blind and at least 55 years old in section 223(d)(1)(B). SSA's regulations, consistent with the Social Security Act, define work disability in 20 CFR 404.1505.

calculation allows for greater weight on a dollar flowing to a low-income SSDI beneficiary, counts the value of time spent out of the labor force, and inflates net government outlays by the excess social cost of raising funds (Gubits et al. 2018a/b).

SSDI and SSI Programs

The SSDI program requires a 10-year history of work of most applicants, and the SSI program requires very low earnings and assets. This means the SSDI and SSI populations differ substantially on average, with SSDI beneficiaries having greater education, incomes, assets, and other attributes associated with better labor market experiences than do SSI recipients (SSA 2020d). Most SSDI applicants have earnings that fall dramatically in the three years prior to application, whereas most SSI applicants have little to no earnings in the year prior to application (Bound, Burkhauser, and Nichols 2003; Costa 2017). SSDI benefits pay a fraction of prior average earnings, which ranges from 90 percent for very low earnings to about 15 percent for higher earnings, with a five-month waiting period after disability onset. In contrast, SSI pays a fixed amount offset by countable income (with slightly more than half of earnings excluded, so the effective marginal tax rate on earnings is slightly less than one-half).

The typical SSDI beneficiary qualifies for Medicare after 24 months, and the typical SSI recipient qualifies for Medicaid immediately. Those dually eligible for both SSDI and SSI, called "concurrent beneficiaries," have their SSI payments reduced once SSDI payments start, and they are eligible for both Medicare (after 24 months) and Medicaid. In most cases, a state pays the Medicare Part B premium for those eligible for SSI and covered by Medicaid. For concurrent beneficiaries, Medicare is the primary payer and Medicaid is the secondary payer.

For adults, the eligibility rules for both SSDI and SSI are the same regarding medical standards, but the programs have different financial eligibility rules. Both SSDI and SSI screen out applicants who work and earn more than the SGA threshold. To qualify for SSDI, an applicant must generally have 40 credits (formerly "quarters of coverage"), 20 of which earned in the last 10 years. (As of 2020, workers earn 1 credit for each $1,410 in wages or self-employment income, to a maximum of 4 credits per year.) Younger workers may qualify with fewer credits. For SSI, an applicant must have countable resources of less than $2,000 for an individual and $3,000 for a couple, which notably excludes the value of a home and one vehicle, among other exclusions.

The medical standards for both programs are strict, with rigorous reviews of medical evidence initially conducted by Disability Determination Service (DDS) agencies in each state in a five-step process. Applicants who pass the financial screen and have a severe impairment could qualify for award if they have a condition(s) on the Listing of Impairments, comprising more than 100 impairments such as cancers, adult brain disorders, and rare disorders that affect children. SSA's Quick Disability

Determination model also quickly identifies diseases and other medical conditions likely to meet SSA's standards so applicants may receive a faster decision.

In that five-step process: For applicants who are financially eligible (step 1) and have a severe impairment (step 2) but do not have a listed condition (step 3), state DDS offices ascertain whether those with severe impairments could work in their past job (step 4) or do other work in the national economy (step 5).

At step 4, the DDS denies applicants whose "residual functional capacity" meets the requirements of past relevant work. At step 5, DDS considers the applicant's remaining capacity, along with other vocational factors—age, education, and work experience—to determine whether the applicant can work in jobs other than that previously held. This determination often involves the use of the Medical-Vocational Guidelines (a set of tables sometimes also known as the "vocational grid") and "medical-vocational profiles."[5]

Over the years, Congress has enacted various rules to incentivize the return or attempted return to work. The Trial Work Period (TWP) allows SSDI beneficiaries to test their ability to work and still be considered disabled. Beneficiaries retain their full disability benefit until their monthly earnings exceed the TWP earnings threshold ($910 in 2020) in at least nine months (not necessarily consecutive) in a rolling 60-month period.

SSA applies various work incentives, such as subsidies or disregarding earnings used to cover impairment-related work expenses, before determining whether net earnings exceed the SGA threshold. When net earnings exceed the SGA threshold after the TWP, SSDI beneficiaries still receive benefits during a three-month Grace Period; but after the Grace Period, SSA suspends benefits in any months of SGA during a 36-month Extended Period of Eligibility (EPE). SSDI beneficiaries who return to work can keep Medicare coverage long after the TWP and EPE have expired—for at least eight and a half years after return to work. SSI recipients may use other work incentives, such as SSA's Plan to Achieve Self–Support; continue to receive Medicaid coverage while working; and have other income exclusions.[6]

In addition to rules supporting return to work, an array of services supports individuals who wish to return to work. SSA's Ticket to Work program supports SSDI beneficiaries and SSI recipients ages 18–64 who want to work by connecting them

---

[5] For more detail on the vocational grid, see the Program Operations Manual System: https://secure.ssa.gov/apps10/poms.nsf/subchapterlist!openview &restricttocategory=04250. For more on medical-vocational profiles, see the Code of Federal Regulations: http://www.socialsecurity.gov/OP_Home/cfr20/404/404-1562.htm and http://www.socialsecurity.gov/OP_Home/cfr20/416/416-0962.htm.

[6] More information on all SSDI and SSI work incentives can be found in the *Red Book* (SSA 2020e) at https://www.ssa.gov/redbook.

with employment service providers including Employment Networks,[7] Vocational Rehabilitation agencies,[8] Work Incentives Planning and Assistance programs,[9] and Protection and Advocacy for Beneficiaries of Social Security organizations.[10] There is a wide range of other private and public return-to-work services across employers and states (Epstein et al. 2020), but little evidence on their impacts (Nichols et al. 2020).

Work and Disability Benefits

Although SSA's demonstrations have addressed multiple issues, the vast majority have dealt with disability and employment. In this section, we highlight the dominant economic theory common to these demonstrations.

Both SSDI beneficiaries and SSI recipients face the possibility of losing benefits if they work, consistent with the intention of the programs to support individuals who are not able to perform SGA. Broadly speaking, SSDI beneficiaries can lose benefits if they earn over the SGA threshold for too long; and SSI recipients lose $1 of benefits for every $2 they earn over $65 in a month. Reductions in benefits are like taxes on earnings—the amount of an individual's income increase is less than the full amount of their earnings, reducing the net gains to work. Receiving the benefits themselves can also reduce their labor market activity by lowering the need to work.

As previously noted, an SSDI beneficiary earning above the SGA level during the EPE can lose all benefits. This "cash cliff" has long been considered a barrier to work because of the potential for losing benefits altogether. Though SSI recipients face a gradual reduction of their payments as they earn more, eventually they, too, lose eligibility. The loss of eligibility is not just for SSI benefits; because most SSI recipients receive Medicaid, they also risk losing access to health insurance. As Livermore, Wittenburg, and Neumark (2014, 3) point out, other barriers include "fear of job failure…lack of job qualifications, lack of reliable transportation, inaccessible or inflexible work environments, and negative employer perceptions of disability."

---

[7] Employment Network (EN)s are private or public organizations that can help with career counseling and assistance with job placement, including helping an individual understand how benefits could be affected by work. This includes ENs that are also part of a state's public workforce system, also referred to as "workforce ENs."

[8] Vocational rehabilitation agencies usually work with individuals who need more substantial services. In some states, this includes intensive training, education, and rehabilitation. Agencies could also provide career counseling, job placement assistance, and counseling on the effect that working could have on Social Security disability benefits.

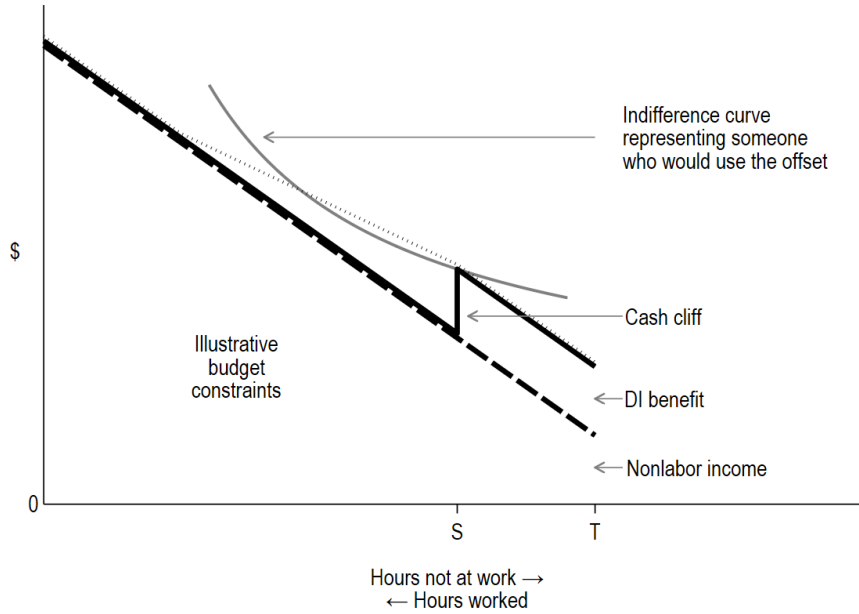
[9] Work incentives counseling is provided by staff who work for the Work Incentives Planning and Assistance programs. Community work incentives coordinators provide counseling about the effect of work on benefits and reach out to beneficiaries who are potentially eligible to participate in federal or state work incentives programs.

[10] Protection and Advocacy for Beneficiaries of Social Security organizations represent eligible beneficiaries to remove barriers to successful employment and will help beneficiaries understand their rights regarding conditions of employment under applicable laws and regulations.

Additionally, many SSDI beneficiaries could simply believe they cannot work because of their impairment or social and familial expectations about what people with disabilities have historically been considered capable of doing.

One way to think about how these demonstrations might function is to consider a simple economic model of behavior (Exhibit 1.2 below). Assume an SSDI beneficiary can earn a given wage per hour, which defines the tradeoff between time spent away from work (on the *x*-axis) and earnings (*y*-axis), along the budget constraint, shown as a thick solid line. During the TWP and Grace Period, the beneficiary's income is their SSDI benefit amount plus their earnings. Once an individual works above the SGA level for nine months, though, they lose all of their SSDI benefit, producing the sudden drop in the thick solid line (the cash cliff).

Exhibit 1.2. Illustrative Work and Earnings Tradeoffs



Note: As hours not at work increase, hours worked decrease, and at T (total number of hours available), hours worked equal zero. At S hours not at work (T−S hours worked), this individual earns exactly the SGA threshold. The first vertical gap at T shows the income level at zero hours worked; that is, nonlabor income. The thick solid line shows a budget constraint for someone eligible for SSDI in the Extended Period of Eligibility (after completing a Trial Work Period). The thick dashed line shows a budget constraint for someone not eligible for SSDI. The thin dotted line shows a budget constraint for someone eligible for SSDI but subject to different rules that impose a benefit offset for earnings above the SGA threshold.

As represented in the exhibit, the beneficiary must choose their hours of work based on this budget constraint and their personal preferences for work, represented by a curved, thin solid line showing all combinations of work and earnings that the

beneficiary values equally—the indifference curve. An individual with these preferences would work to earn exactly at the SGA threshold when eligible for SSDI, and not more when subject to the cash cliff. They might, however, work more with a benefit offset above the SGA threshold (a budget constraint with the offset is shown as the dotted line in Exhibit 1.2); this benefit offset, reduces benefits slowly above the SGA threshold, rather than zeroing them out.

In economic terms, disability benefits will decrease hours worked, due to both "income effects" and "substitution effects." The income effect refers to a reduction in work when workers are slightly richer and therefore desire slightly more time away from work, whereas the substitution effect refers to a small reduction in the relative benefit of working. But disability benefits are not small changes to the budget constraint, and they produce very different nonlinear changes. That is, the budget constraint looks quite different in shape for someone eligible for disability benefits (the solid line in Exhibit 1.2, rather than the dashed line for someone not eligible for benefits). Conceptually, we can measure both income and substitution effects of providing disability benefits if we limit our attention to the EPE and adopt a very generous interpretation of smoothed budget constraints.

There are many areas where policies, programs, and services can influence the barriers to, and net benefits of, working. The SSA demonstrations reviewed in this volume have tested a variety of interventions. Counseling or information campaigns can change an individual's expectations about work, or preferences for work. Employment-related services can reduce the barriers to accessing a job, better prepare an individual to self-advocate, identify appropriate and reasonable accommodations, or otherwise support employment. Program rules can be modified to reduce, delay, or smooth the perceived negative consequences on benefits of work or reduce any administrative burdens related to working while receiving benefits (e.g., reporting requirements). Interventions can operate via direct subsidies, rule changes, employers, or other ways to produce changes in labor demand.

Several SSA demonstrations have explicitly attempted to change the budget constraint by offsetting SSDI or SSI benefits by a smaller amount, so that an individual never sees a drop in total income due to work. The most prominent example is a benefit offset (the thin dotted line in Exhibit 1.2) of $1 lower SSDI benefit for every $2 in earnings above the SGA threshold, but other demonstrations included starting to offset benefits at a lower earnings level or altering the offset in SSI rules. Demonstrations have also shifted the budget constraint by providing services, reducing the costs of services, or supplementing wages—allowing the beneficiary to effectively keep more income for any given level of work.

Some counseling services provided in the demonstrations serve to change the information available to a beneficiary or recipient, often specifically focusing on explaining the incentives inherent in the budget constraint and the various work incentives available to them. The theory of change embodied in this type of intervention is that an SSDI beneficiary may face a budget constraint such as in

Exhibit 1.2 but not fully understand it. Survey results in the demonstrations support the idea that many participants do not fully understand the current-law rules they are subject to, suggesting a role for information provision.

Still other demonstrations have approached the issue by attempting to change the participant's perceptions of work, by providing information and counseling about the advantages of work. A change in perceptions or preferences favoring work might involve "lowering the disutility of work" (as an economist conceives of utility). This would then change where an individual locates—or believes they can locate—on the budget constraint line or even change the utility curves themselves. Compared with modest changes in the budget constraint, modest changes in preferences can, potentially, produce very large changes in observed behavior. There is very little research on what forces act to produce the type of indifference curve seen in Exhibit 1.2, but society at large and media representations, local communities, habits formed from past experience, and family members may play important roles.

Other federal programs and policies try to address similar issues. For example, the Earned Income Tax Credit changes a worker's budget constraint by subsidizing work, which tends to increase hours worked (Nichols and Rothstein 2016). The Administration for Children and Families within the US Department of Health and Human Services has created a Pathways to Work Evidence Clearinghouse[11] with information on programs designed to support low-income workers succeeding in the labor market. Two other clearinghouses covering relevant interventions[12] are the US Department of Labor's Clearinghouse for Labor Evaluation and Research[13] and the US Department of Education's What Works Clearinghouse,[14] the latter focused on youth.

SSA'S DEMONSTRATION AUTHORITIES

To provide a common understanding of what SSA can and cannot do in its demonstrations, this section provides details about SSA's relevant statutory authorities.

History of Authorities

Congress first authorized SSDI and SSI demonstrations related in Section 505 of the Social Security Disability Amendments of 1980. Section 505 specifically created a new, permanent Subsection 1110(b) of the Social Security Act allowing SSA to test

---

[11] See https://pathwaystowork.acf.hhs.gov/ (accessed May 30, 2021, at which time it showed 176 interventions in 244 studies reviewed).

[12] SSA's Interventional Cooperative Agreement Program (ICAP) prioritizes evidence from these three clearinghouses; more information on ICAP can be found at grants.gov under Opportunity Number ICAP-ICAP-21-001.

[13] See https://clear.dol.gov/ (accessed May 30, 2021).

[14] See https://ies.ed.gov/ncee/wwc/ (accessed May 30, 2021).

changes of *SSI* program rules ("waivers"), one kind of intervention that can be tested in a demonstration. This waiver authority complemented the existing general research authority provided by Subsection 1110(a). Section 505 also authorized SSA to test changes to *SSDI* program rules, but because it did not create a new permanent authority (as Subsection 1110(b) had), this provision sunset after five years. Throughout the 1980s and 1990s, Section 505 was renewed for various periods (with some lapses), until Congress created a specific SSDI demonstration authority as part of the Ticket Act.

The Ticket Act contained two demonstration provisions. Section 301 of the Ticket Act created Section 234 of the Social Security Act, which very closely mirrored the language of Section 505 of the 1980 Amendments. Section 302 of the Ticket Act directed SSA to conduct a $1 for $2 benefit offset demonstration (which eventually became BOND).[15] Section 234 initially sunset on December 17, 2004, after which SSA could not initiate any new SSDI demonstrations or continue existing ones. The Social Security Protection Act of 2004 extended demonstration authority through December 18, 2005, and allowed SSDI demonstrations initiated by December 17, 2005, to be completed thereafter. Once those projects ended, all SSDI demonstration activity stopped (with the notable exception of BOND).

Congress last renewed Section 234 as part of the Bipartisan Budget Act of 2015, allowing SSA to begin SSDI demonstrations until December 31, 2021, and requiring these demonstrations to end no later than December 31, 2022. This tied closely to the projections at the time of when the Disability Insurance Trust Fund would become insolvent. In Section 823 of that Act, Congress also instructed SSA to conduct a second $1 for $2 offset demonstration (the Promoting Opportunity Demonstration, or POD), specifically creating Section 234(f) of the Social Security Act.

Specific Rules and Changes over Time

Section 234 and Section 1110(b) each have specific rules for how SSA can conduct demonstrations. There is some commonality between the two, but also distinctions. One important distinction is how SSA funds projects. SSA requests an apportionment directly from the Disability Insurance Trust Fund for projects authorized under Section 234. Projects authorized under Sections 1110(a) and (b) are instead funded from the general Treasury and are included in SSA's annual budget request to Congress.

In addition to requiring that a project be related to the SSDI program (or the Ticket to Work program), Section 234 specifies that demonstrations be about changing the SSDI program in some way. They can cover topics related to alternative treatments of work activity and other rules, such as changing the 24-month waiting period for

[15] As a result of this, BOND is technically authorized by Section 302 of the Ticket Act and *not* Section 234, although their general reporting, funding, and other provisions closely mirror each other.

Medicare. Projects conducted under Section 234 are required to "be of sufficient scope and carried out on a wide enough scale" to allow SSA to adequately test the policy while ensuring that "the results derived…will obtain generally in the operation of the disability insurance program…without committing such program to the adoption of any particular system either locally or nationally." Section 234 also requires that projects be "expected to yield statistically significant results." Although this does not require nationally representative participation in a statistical sense, it does mean that the policies tested should be relevant to sufficiently broad situations to inform national policy.

In the Bipartisan Budget Act of 2015, Congress included specific language that limited the scope of demonstrations under Section 234. Congress inserted specific language that such projects be "designed to promote attachment to the labor force." Additionally, SSA had previously signed onto the federal Common Rule for human subjects protections,[16] which includes Institutional Review Board review and limits instances where there could be compulsory participation. Even so, Congress inserted language in the Social Security Act specifically requiring that SSDI demonstration participation be voluntary, including revocable informed written consent. This effectively restricts SSA to testing only policies that are more generous to the beneficiary or recipient.[17] Our review of demonstrations for this volume leads us to believe SSA is meeting a substantially higher ethical standard than required. To protect participants from harm, SSA abides by the Common Rule and other procedures. But equitable evaluation and formulation of the causal models to be tested also call for a focus on equity and inclusion. We anticipate this will be a priority for SSA moving forward.

Section 234 also has specific reporting requirements. Ninety days prior to implementing a demonstration, SSA must notify Congress of its plans for the demonstration. SSA must also provide Congress with an annual report by September 30 on demonstrations covered by Section 234, as well as final reports to Congress 90 days after a project ends.

As noted, Section 1110 has two parts. Section 1110(a) is a general research authority allowing SSA to enter into contracts, grants, and jointly financed cooperative agreements. These projects cover a large swath of topics

> relating to the prevention and reduction of dependency, or which will aid in effecting coordination of planning between private and public welfare agencies or which will help improve the

[16] At least 16 federal departments and agencies have issued final revisions to the Federal Policy for the Protection of Human Subjects (known as the Common Rule). A revised final rule was published in the Federal Register on January 19, 2017, on pages 7149 to 7274 (https://www.govinfo.gov/content/pkg/FR-2017-01-19/pdf/2017-01058.pdf).

[17] An exception is POD, under which some beneficiaries could be made worse off; however, as noted, Congress mandated that SSA conduct POD.

> administration and effectiveness of programs carried on or assisted under the Social Security Act and programs related thereto.

This is the broad authority that SSA has used to conduct projects related to early intervention and other topics that do not require changes to program rules.

Section 1110(b) allows SSA to waive program rules related to the SSI program. Like the current version of Section 234, under Section 1110(b), when SSA wants to waive program rules, it must obtain revocable informed written consent. Projects under Section 1110(b) must also not "result in a substantial reduction in any individual's total income and resources as a result of his or her participation in the project."

Although Section 1110 does not have any specific notification or reporting requirements (the budget process alerts Congress to what SSA is planning to do), SSA cannot enter into contracts or jointly financed cooperative arrangements without obtaining "the advice and recommendations of specialists who are competent to evaluate the proposed projects." SSA typically satisfies this requirement by holding technical expert panels before awarding new contracts; for jointly financed cooperative arrangements, the award process includes reviews by experts.

Recent Legislative Proposals

SSA's Fiscal Year (FY) 2018 through FY 2021 budgets included a legislative proposal to extend Section 234 and *require* participation of recipients and beneficiaries, when appropriate. As previously alluded to, many policy proposals, such as time-limited benefits or triage systems suggested in policy circles, could be very difficult to recruit for, limiting the usefulness of results based on voluntary studies. Mandatory participation would allow SSA to test these types of interventions in implementable demonstrations.

As part of that legislative proposal, a new expert panel would be established to identify changes to program rules based on the results of successful demonstrations and other evidence. These changes would be expected to reduce SSDI and SSI outlays by 1 percent as of 6 years after the new authority is passed, increasing to 5 percent after 10 years. Savings of that level are unlikely without policies substantially different from current law, if the cost-benefit analyses and findings of SSA's existing demonstrations hold generally.

As discussed later in this chapter, to achieve a 5 percent savings in total program cost, SSA would likely have to make changes outside the scope of anything allowed under current law or anything tested in prior demonstrations. Whether such changes would be attractive to volunteers is uncertain, and evaluations to test such changes would likely be difficult.

SSA's more-recent legislative proposal in its budget does not request the ability to conduct mandatory tests, requesting instead a simple extension of the existing authority.

Recent Developments

In January 2019, the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act) became law. Among other requirements, the Evidence Act requires agencies to increase their use and documentation of evidence-building methods and approaches. Demonstrations, such as those conducted by SSA and summarized in this volume, address many of the requirements of the Evidence Act. They are rigorous tests of new policies, services, supports, procedures, and the like, intended to inform policymakers whether the change has the intended effects.

As such, even if the specific policies tested by demonstrations are found to not work as intended—especially then—demonstrations serve a valuable role in the evidence-building process. SSA's demonstration experience positions it to be a fruitful partner to other federal agencies developing these capabilities.

SSA'S DEMONSTRATIONS

In this section, we provide an overview of SSA's many demonstrations. The remaining chapters in this volume delve deeper into their cross-cutting themes and lessons.

Overview

In the 1980s, SSA tested the effectiveness of transitional employment as a means of helping SSI recipients with intellectual disability become more self-sufficient, in the Transitional Employment Training Demonstration (TETD). In the 1990s, SSA tested in Project NetWork whether different forms of outreach and case management increased participation in Vocational Rehabilitation services.

Since the early 2000s, SSA has completed many new demonstrations. SSA conducted the Accelerated Benefits (AB) demonstration to test whether earlier access to health care improved employment among new SSDI beneficiaries. SSA conducted the Youth Transition Demonstration (YTD) to test whether providing employment services and other supports to youth receiving or potentially eligible to receive benefits improves self-sufficiency. SSA conducted the Benefit Offset Pilot Demonstration (BOPD) and BOND to test whether alternative benefit structures increase employment among SSDI beneficiaries. SSA also conducted the Mental Health Treatment Study (MHTS) to test whether supports to beneficiaries with mental impairments improve their employment outcomes.

Currently, SSA is conducting POD to test the effect of additional changes to the SSDI structure on beneficiary employment and benefits; the Supported Employment Demonstration (SED) to determine whether providing services and supports to denied applicants reduces the need for future benefits; and the Promoting Readiness of Minors in SSI (PROMISE) demonstration, in concert with the Department of Education, to test whether family-based supports improve adult employment outcomes in families

with child SSI recipients.[18] In each of these demonstrations, SSA has engaged contractors to help with implementation and evaluation.

SSA also conducts demonstrations without contractors, often with the help of government partners. In the 2010s, SSA partnered with local government agencies in the Homeless with Schizophrenia Presumptive Disability (HSPD) Pilot demonstration to test whether presumptive SSI payments improve the application process for a hard-to-serve homeless population. SSA also conducted several pilot mailing studies with the support of the Social and Behavioral Sciences Team, formerly in the White House and now the Office of Evaluation Sciences at the General Services Administration.

SSA's recent demonstrations typically involve both contractors and government partners. SSA's ongoing Retaining Employment and Talent after Injury/Illness Network (RETAIN) demonstration is a joint project with the Department of Labor to help individuals with recent impairments remain in the labor force. Another new project is the Promoting Work through Early Interventions Project (PWEIP) with the Administration for Children and Families to identify ways to support the self-sufficiency of low-income individuals. SSA also recently convened technical expert panels to help provide feedback on selected ideas for future demonstrations.[19]

Aside from SSA's demonstrations, many other federal government research projects have generated evidence on SSDI and SSI benefits. For example, the Structured Training and Employment Transitional Services (STETS) demonstration, sponsored by the Department of Labor, produced findings comparable to TETD. More recently, the Demonstration to Maintain Independence and Employment (DMIE), sponsored by the Department of Health and Human Services, produced findings comparable to SSA's AB demonstration.[20] A variety of welfare-to-work experiments in the 1990s examined a population overlapping the SSI recipient pool. Many evaluations of labor market policies (Card, Kluve, and Weber 2010; Klerman 2020), including those related to Unemployment Insurance (Klerman 2020; Meyer 1995), have relevant findings. In social policy research, evaluations that are part of projects—such as Building Evidence on Employment Strategies for Low-Income Families and the Next Generation of Enhanced Employment Strategies (Martinson et al. 2021), both

---

[18] See www.ssa.gov/disabilityresearch for more information on existing SSA demonstrations. SSA also produces an annual Section 234 report on demonstrations authorized under one of its two demonstration authorities (Section 234 of the Social Security Act) that includes a summary of findings and papers based on those demonstrations. The most recent version of this report can be found at https://www.ssa.gov/legislation/other.html.

[19] See https://www.ssa.gov/disabilityresearch/demos.htm for these technical expert panel reports.

[20] The DMIE, authorized under the Ticket Act, included random assignment demonstrations in Hawaii, Kansas, Minnesota, and Texas. The intervention paired comprehensive health insurance coverage with employment supports to help to maintain participants' employment, health, and independence from public assistance. The states could provide health insurance coverage that was equivalent to their standard Medicaid benefit package or "wraparound" coverage that supplements public or employer-sponsored coverage.

run by the Administration for Children and Families with support from SSA as part of PWEIP—will have relevant findings, as well.

Evaluating Demonstrations

Evaluating policy changes in disability programs has proven challenging because of the myriad programs that serve individuals and numerous agencies involved in their administration. Wittenburg, Mann, and Thompkins (2013, 4) claim

> [E]ach individual program plods along, trying to improve its part of the overall system in ways that add up to very little overall progress. In reviewing evaluations of 27 federally sponsored employment programs, policies, and initiatives conducted since 2000, Livermore and Goodman (2009) found that many were not rigorously evaluated due, in part, to their limited focus and lack of a planned evaluation framework.

There are other pressures that lead to evaluation challenges; for example, the design of POD was mandated by Congress, with elements that made the evaluation design challenging.[21]

There is a large body of related quasi-experimental academic literature. Gelber, Moore, and Strand (2017), for example, estimate that the change in slope at the upper bend point of the SSDI benefit formula implies that if SSDI payments were increased by $1, beneficiaries would decrease their earnings by about $0.20. The authors conclude that most of the drop in earnings associated with receipt of disability benefits (measured in Bound 1989; Bound, Burkhauser, and Nichols 2003; French and Song 2014; Maestas, Mullen, and Strand 2013; von Wachter, Song, and Manchester 2011) is due to income effects. Another example, using a difference-in-differences approach (Mullen and Rennane 2017), has several possible interpretations of the relative importance of income and substitution effects. As important as this distinction is to disability policy, it seems we do not have decisive evidence. But what evidence we have points to substitution effects not driving much of the low employment rates among those receiving disability benefits.

---

[21] The legislation authorizing POD required a demonstration that would be externally valid; that is, results would apply generally to the operation of the disability program. But the study was required also to use only volunteers, advise them in detail of their incentives, and allow volunteers to leave the demonstration at any time. Because some volunteers who earned between TWP and SGA would be worse off, those induced to earn between TWP and SGA amounts can be expected to be differentially underrepresented in the treatment group, compromising external validity. Per Hock et al. (2020), "the evidence suggests that a key motivation for POD withdrawals to date is the potential for POD to reduce income compared to current SSDI rules, but early withdrawal rates were not high enough to make this a concern for the impact analysis." Wiseman (2016) highlights some of the design challenges inherent in the legislative authority for POD.

SSA has tended to rely on experimental studies to isolate the effect of interventions. The SSA demonstrations all had distinct target populations and involved multiple partnerships, as depicted in Exhibit 1.3. The interventions were of varying types; the findings, therefore, varied across demonstrations, settings, and populations. (The next section describes key findings on a more comparable scale.[22])

Exhibit 1.3. Overview of Prominent SSA Demonstrations, by Initial Year

Demonstration Name and Date	Design Features	Findings
Transitional Employment Training Demonstration (TETD), 1985–1987	Randomly assigned SSI recipients ages 18–40 with intellectual disability who volunteered to potentially permanent competitive jobs and specialized on-the-job training and supports for 1 year	Increased monthly earnings $64, and decreased monthly SSI payments $7, on average, after 3 years (Thornton and Decker 1989; Thornton, Dunstan, and Schore 1988)
Project NetWork, 1992–1996	Randomly assigned SSDI beneficiaries and SSI applicants and recipients who volunteered to case management, employment, and rehabilitation services	Increased average annual earnings by $220 per year over 2 years, but no discernable impact on benefit receipt (Kornfeld and Rupp 2000; Kornfeld et al. 1999; Rupp, Wood, and Bell 1996)
State Partnership Initiative (SPI), 2003–2005	Included 12 projects funded by SSA (of which only 10 reported impact estimates and only NH, NY, and OK reported usable estimates using random assignment) that gave SSDI beneficiaries and SSI recipients information about the effect of work on benefit receipt and work incentives, and modified program rules to allow SSI recipients to earn and save more	Either no effect or a negative and statistically significant effect on annual earnings (Peikes et al. 2005)
Benefit Offset Pilot Demonstration (BOPD), 2005–2006	Randomly assigned SSDI beneficiaries in CT, UT, VT, and WI who volunteered to receive a "benefit offset" in SSDI benefit payments rather than their payment stopping when earnings exceed SGA	Earnings indistinguishable but benefit payments about 5% higher in the treatment group; treatment group had about 4% more beneficiaries with earnings above annualized SGA (Weathers and Hemmeter 2011)

[22] We do not include in the next section demonstrations without publicly available evaluation findings, such as the SSI/SSDI Outreach, Access, and Recovery (SOAR) demonstration started by SSA in Baltimore in 1993. SOAR aims to increase the award rate and reduce the time from application to decision for adults who are homeless or at risk of being homeless and have a mental illness, medical impairment, and/or a co-occurring substance abuse disorder. It is comparable to the HSPD Pilot and to PWEIP's projects BEES and NexGen. The SOAR program is now funded by the Substance Abuse and Mental Health Services Administration (SAMHSA, n.d.).

Demonstration Name and Date	Design Features	Findings
Mental Health Treatment Study (MHTS), 2006–2010	Randomly assigned SSDI beneficiaries with schizophrenia or an affective disorder to supported employment services and systematic medication management services. Some disincentives removed	Increased 24-month employment rate and earnings, improvement in mental health status and quality of life, but slight decline in physical health status; no detectable difference in earnings above the SGA threshold (Frey et al. 2011)
Youth Transition Demonstration (YTD), 2006–2012	Randomly assigned youth ages 14–25 receiving SSDI or SSI disability benefits, or at high risk of receiving benefits, to receive various work-promoting services and incentives, including work experiences, youth empowerment, family support, system linkages, social and health services	No detectable overall impact on earnings; increased employment about 4%, and disability benefits more than $500 a year higher, at the end of 3 years (Fraker, Mamun, et al. 2014)
Accelerated Benefits (AB), 2010–2015	Random assignment of new SSDI beneficiaries who volunteered and had no health insurance to comprehensive health insurance or to insurance plus medical care management and access to employment and benefits counseling	No detectable impact on employment in year one (Michalopoulos et al. 2011); higher employment and earnings in year two, but no detectable impacts on employment and earnings in year three (Weathers and Bailey 2014)
Benefit Offset National Demonstration (BOND), 2011–2016	Randomly assigned SSDI beneficiaries to receive benefit offset; Stage 1 included mandatory participation, and Stage 2 included volunteers only	No detectable impact on average earnings, 1% higher average benefits among treatment group in Stage 1 and 4% higher in Stage 2; more beneficiaries earned above annualized SGA threshold (Gubits et al. 2018a/b)
Homeless with Schizophrenia Presumptive Disability (HSPD) Pilot, 2012–2014	Compared SSI-eligible individuals with schizophrenia or schizoaffective disorder who got application assistance and presumptive disability SSI payments versus quasi-experimental comparison group that did not	Time between the SSI claim and first SSI payment shortened 3–5 months; higher initial allowance rate and payment status (Bailey, Goetz Engler, and Hemmeter 2016)

Demonstration Name and Date	Design Features	Findings
Promoting Readiness of Minors in SSI (PROMISE), 2016–2018	Randomly assigned SSI recipients ages 14–16 to different intensive case management and connection to community resources (e.g., benefits counseling and financial education, different types of career exploration and work-based learning experiences, promotion of self-esteem and self-advocacy, and parent training and information sessions)	Increased youth receipt of transition services, youth paid employment, family member receipt of support services during the first 18 months after enrollment, and youth receipt of job-related training or credentials. In four states (AR, CA, MD, and WI), increased youth total income from earnings and SSA payments; only in CA, reduced youth receipt of any SSA payments (Mamun et al. 2019)
Supported Employment Demonstration (SED), 2017–present	Randomly assigned denied applicants alleging a mental impairment to receive Individual Placement and Support (employment services integrated with behavioral health and other services)	No evaluation results yet
Promoting Opportunity Demonstration (POD), 2017–present	Randomly assigned SSDI beneficiaries who volunteered to benefit adjustment (offset starts at lower income than in BOPD and BOND)	No discernable impacts on earnings, employment, or benefits (Mamun et al. 2021)
Promoting Work through Early Interventions Project (PWEIP), 2017–present	Supports two existing Administration for Children and Families projects: Building Evidence on Employment Strategies for Low-Income Families (BEES) and Next Generation of Enhanced Employment Strategies (NextGen)	No evaluation results yet (Martinson et al. 2021); see also Chapter 5 in this volume
Retaining Employment and Talent after Injury/ Illness Network (RETAIN), 2018–present[a]	Intervention (case management and connection to occupational health services), target population, and evaluation method vary by state (CA, CT, KS, KY, MN, OH, VT, WA)	No evaluation results yet

Source: SSA (2020a) and works cited in the exhibit.

Note: This table includes only demonstrations described in publications that include evaluation results (or plans for evaluation) where outcomes included benefits, earnings, and/or employment rates; there are other demonstrations that do not meet this criterion, including two described in the Appendix: Benefits Entitlement Services Team (BEST) and Homeless Outreach Projects and Evaluation (HOPE).

[a] Ongoing as of publication (2021).

[b] Evaluation to be implemented in Phase 2. SSA (2020a) notes that "due to the impacts of the COVID-19 pandemic, DOL…delayed the publication of the Phase 2 Funding Opportunity Announcement until FY 2021" and "the evaluation contractor will produce an interim impact report in late FY 2025 and the final evaluation impact report in FY 2026."

A wide variety of non-SSA demonstrations that promoted training or work could be relevant to people who might apply for disability benefits. For example, the National Supported Work demonstration aimed at long-term welfare recipients (Hollister, Kemper, and Maynard 1984), and the National Job Training Partnership Act Study was designed to served economically disadvantaged adults and out-of-school youth (Bloom et al. 1997). The Health Profession Opportunity Grants (HPOG) and the Pathways for Advancing Careers and Education (PACE) interventions aim to improve employment for low-skilled adults (Gardiner and Juras 2019; Peck et al. 2019).

A handful of demonstrations are even more directly related to improving employment among persons with disabilities. The Structured Training and Employment Transitional Services (STETS) and the Demonstration to Maintain Independence and Employment (DMIE) are depicted in Exhibit 1.4 and subsequent exhibits, but there are many comparable demonstrations and experiments worldwide. There are also many state and local return-to-work initiatives (Nichols et al. 2020).

Exhibit 1.4. Overview of Related Non-SSA Demonstrations

Demonstration Name and Date	Design Features	Findings
Structured Training and Employment Transitional Services (STETS), 1981–1983	Two-armed random assignment in five sites (Cincinnati, OH; Los Angeles, CA; New York, NY; St. Paul, MN; and Tucson, AZ) for referred individuals with intellectual disability provided transitional employment building on the National Supported Work design: initial training and support, placement in on-the-job training, and withdrawal of support with follow-up services	Increased employment (31% vs. 19%) and earnings ($36 vs. $21 per week) at month 22 (Kerachsky et al. 1985)
Demonstration to Maintain Independence and Employment (DMIE), 2007–2009	Random assignment of population that varied by state (HI working adults ages 18–62 with diabetes; KS working adults ages 18–64 with various conditions; MN working adults ages 18–60 with serious mental illness; TX working adults ages 21–60 with either severe mental illness or behavioral health diagnoses) to receive services that varied by state (including case management, health coverage, and employment services)	KS and MN had modestly increased employment, but HI and TX did not; none discernably affected average earnings (Whalen et al. 2012)

Source: Works cited in the exhibit.

Average Findings on Employment and Benefits

The general findings in the demonstrations in Exhibits 1.3 and 1.4 above suggest relatively modest impacts. As a means of systematically assessing the state of the evidence, we conducted a meta-analysis of findings from these evaluations, reporting the results in Exhibits 1.5, 1.6, and 1.7 below. We order the demonstrations in the exhibits by the primary year of earnings data collection, to capture variation across the business cycle, but report dollar values in 2020 dollars. A diamond-shaped symbol at the bottom of each figure captures the "Overall" average effect across all those reported. Across these many demonstrations:

- Exhibit 1.5 shows the average effect on benefits is +$72 per year (with a confidence interval from 37 to 107) in 2020 dollars, but the heterogeneity in impacts suggests there is not a single common underlying effect across studies.
- Exhibit 1.6 shows the average effect on earnings is +$97 per year (with a confidence interval from 41 to 153); again, there is substantial variation across studies. The various effects on earnings are quite small, aside from the early STETS and TETD studies.
- Exhibit 1.7 shows the average effect on employment is +1.7 percentage points (with a confidence interval from 0.9 to 2.5). This is a relatively small average effect, but with some notable outliers; for example, MHTS, STETS, TETD, and some of the youth-focused demonstrations in PROMISE and YTD achieved meaningful increases in employment rates.

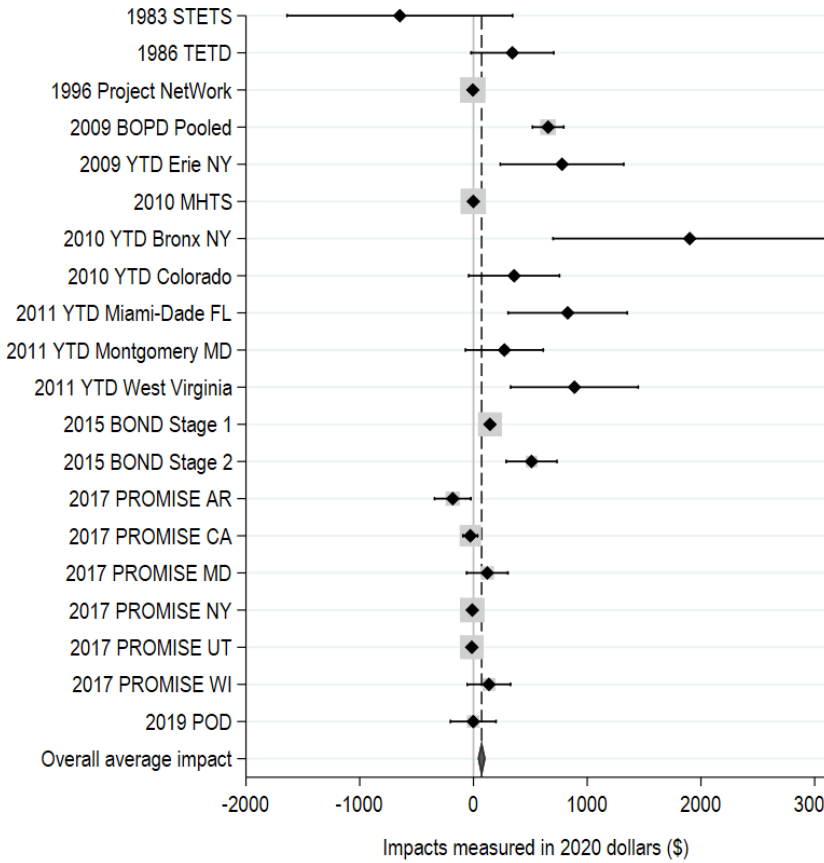
We have made some simplifying assumptions to put these various demonstrations on the same graphs, and different analysts could come up with a different overall average for each synthesis; but we are confident the overall averages would be qualitatively similar across alternative approaches. We do not suggest that there is any obvious pattern to be seen in the larger or smaller estimates.

These syntheses suggest one clear takeaway. The findings reflected in these three exhibits lead to a general conclusion about what the field has learned about disability policy from the demonstrations we have examined: *It is much easier to increase employment than it is to increase net earnings appreciably or to reduce disability benefits.* Indeed, a 1.7 percentage point average increase in employment represents a meaningful increase for this population, where employment rates are relatively low. The corresponding implications of that employment gain for earnings and benefits—less than $100 over the course of the year—seem less life changing.

That said, there are non-monetary benefits to work that have personal and social value. For example, about 45 percent of beneficiaries and recipients reported in 2015 that their goals included working or advancing in their careers or that they saw themselves working in the near future (SSA 2020d). Even without much of an increase in earnings, greater employment could reflect improvements in well-being for some

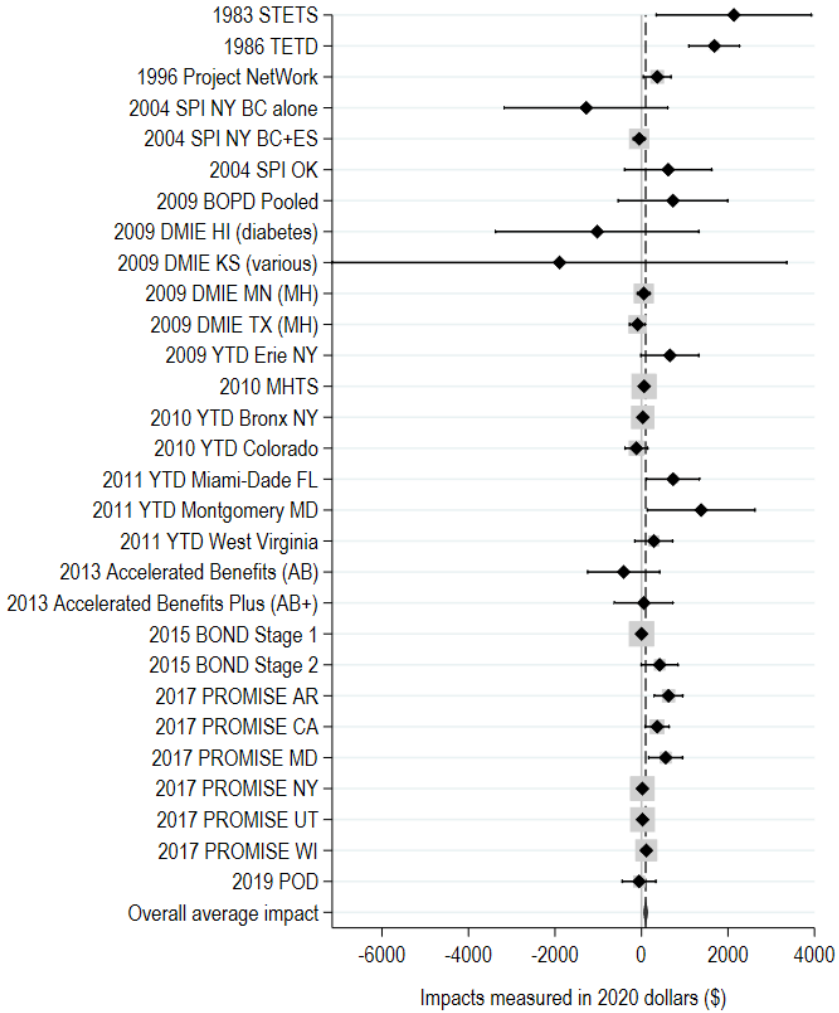
SSDI beneficiaries and SSI recipients. Moreover, as we noted above, the average values from the meta-analysis mask substantial variation in impacts. Not only is variation in the impacts wide, but importantly, the populations and purposes of the demonstrations also vary widely. These simple analyses do not capture the full range of questions relevant to policymakers. The remainder of this volume helps fill these gaps.

Exhibit 1.5. Average of Effects on Annual Benefits across Evaluations in Demonstrations



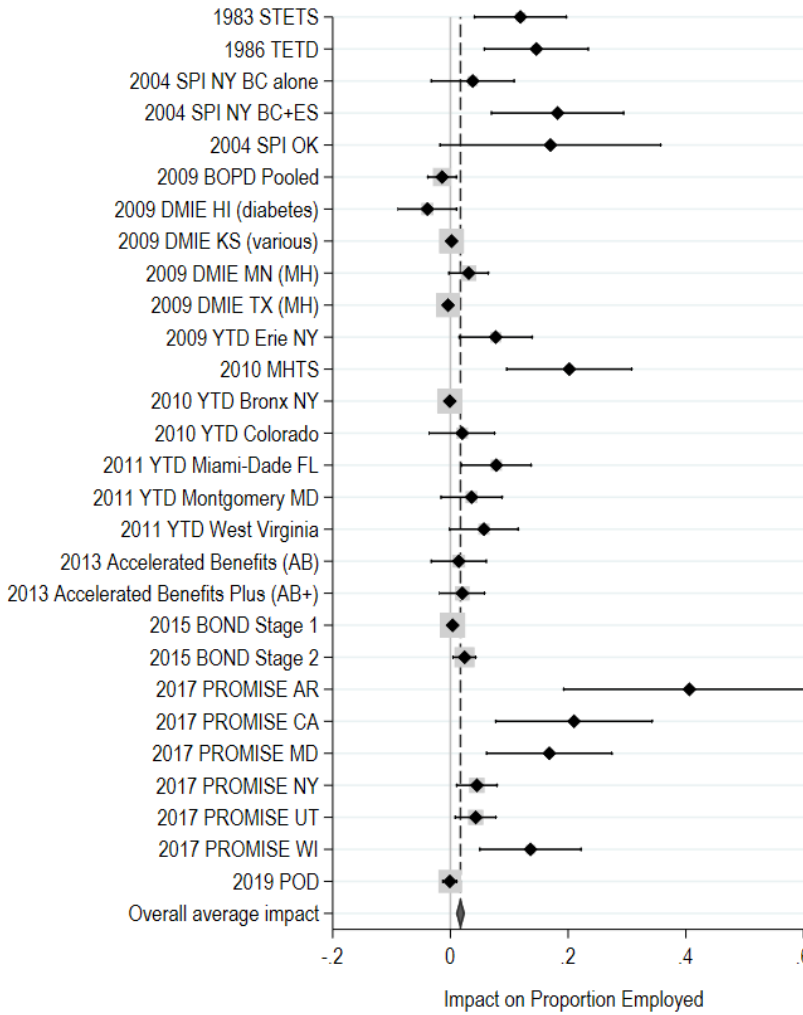
Source: Authors' computations from individual evaluations.
Note: All dollar values in inflation-adjusted 2020 dollars. In the exhibit, the more precisely estimated impacts have greater weight in the average, shown as larger boxes centered on the point estimates. The less precise estimates have wide confidence intervals and appear on either side of the zero line. The overall average impact and its confidence interval appear at the bottom. A measure of heterogeneity, I^2 (Higgins and Thompson 2004), shows that 92 percent of the variation is attributable to heterogeneity across studies. This heterogeneity suggests we should not interpret the average of +$72 per year (with a confidence interval from 37 to 107) as being the common effect across studies.

Exhibit 1.6. Average of Effects on Annual Earnings across Evaluations in Demonstrations



Source: Authors' computations from individual evaluations.
Note: All dollar values in inflation-adjusted 2020 dollars. In the exhibit, the more precisely estimated impacts have greater weight in the average, shown as larger boxes centered on the point estimates. The less precise estimates have wide confidence intervals and appear on either side of the zero line. The overall average impact and its confidence interval appear at the bottom. A measure of heterogeneity, I^2 (Higgins and Thompson 2004), shows that 73 percent of the variation is attributable to heterogeneity across studies. This heterogeneity suggests we should not interpret the average of +$97 per year (with a confidence interval from 41 to 153) as being the common effect across studies.

Exhibit 1.7. Average of Effects on Employment Rates across Evaluations in Demonstrations



Source: Authors' computations from individual evaluations.
Note: In the exhibit, the more precisely estimated impacts have greater weight in the average, shown as larger boxes centered on the point estimates. The less precise estimates have wide confidence intervals and appear on either side of the zero line. The overall average impact and its confidence interval appear at the bottom. A measure of heterogeneity, I^2 (Higgins and Thompson 2004) shows that 81 percent of the variation is attributable to heterogeneity across studies. This heterogeneity suggests we should not interpret the average of +1.7 percentage points (with a confidence interval from 0.9 to 2.5) as being the common effect across studies.

CHAPTER OVERVIEWS

To guide readers through the rest of this volume, in this section we provide brief overviews of all the subsequent chapters. We also provide a summary of selected lessons. Each chapter has a short introduction then discusses relevant history, policy setting, and current program rules. Their next sections discuss topic-specific relevant theory, summarize relevant empirical evidence, and then explore lessons both for policy and for SSA's future learning agenda.

Chapters 2 and 3 focus on the methodological aspects of the demonstrations. In Chapter 2, Burt Barnow and David Greenberg review the design of evaluations that are typically part of a demonstration. They review the designs of past demonstrations' evaluations and discuss the implications for what questions those designs can address. They also offer thoughts on alternative designs that SSA might consider in the future for greater learning. They encourage SSA to consider multiple, varied treatment arms and factorial designs to help determine the role of the component parts of the interventions. Additionally, Barnow and Greenberg suggest additional designs—for example, "stepped-wedge"—that could help provide additional estimates. They also suggest going beyond the intent-to-treat (ITT) estimates that have been the default in SSA's existing demonstrations and encourage treatment-on-the-treated (TOT) estimates, as well. Finally, they note how important process analyses are and suggest that an increased use of fidelity measures could help SSA learn more about the specific interventions.

In Chapter 3, Robert Weathers and Austin Nichols discuss ways to improve the use of evaluation findings and implications for how demonstrations should be used. They focus on questions policymakers have historically wanted answered and how to better communicate findings to meet those needs. Weathers and Nichols encourage strong theoretical models (including logic models) to underpin the demonstrations and clarify their goals. While acknowledging the existence of tradeoffs, they also urge SSA to consider going beyond the single-intervention tests by looking at broad ranges of similar policy options. Looking inside the "black box," through multiple treatment arms, factorial designs, etc., is important to answering questions related to why something worked. The authors also encourage additional uses of qualitative findings and reanalyzing the data from past demonstrations to extend the analyses delivered in evaluation contract reports.

Chapter 4 reviews the lessons from SSA's return-to-work demonstrations such as BOND. Jesse Gregory and Robert Moffitt describe the incentives that individuals receiving SSDI and SSI benefits face. They describe the history of efforts to improve this population's work outcomes, including how the demonstrations align with economic theory and what researchers expected to find. They also propose several considerations for SSA's next generation of demonstrations, including ideas related to Chapters 2 and 3. One important lesson they note is that most efforts to increase employment, earnings, and labor force engagement (i.e., working or looking for work) do not have large effects. As a result, it may be necessary to reconsider expectations

about how many beneficiaries and recipients will go back to work. Promising areas to explore could include work incentives such as the Earned Income Tax Credit that provide additional income above and beyond the current benefit (as opposed to just not taking as much away) at first.

Next, Kevin Hollenbeck in Chapter 5 provides an overview of SSA's efforts to explore policies related to individuals not yet receiving disability benefits. This includes early interventions that might prevent some of them from needing the SSDI or SSI programs for support. Programs that effectively reduce the need for SSDI or SSI could both improve individuals' economic well-being and reduce government expenditures on these programs, allowing for a more efficient use of program resources. Hollenbeck shares lessons from his review of both US and international programs related to SSDI and SSI. He notes that Individual Placement and Support has been tested in several settings, and though it does appear to have some success in improving labor market outcomes, there is little evidence that this translates into reductions in SSI or SSDI benefits. He recommends testing interventions targeting older denied applicants rather than a large swath of potential applicants.

In Chapter 6, David Wittenburg and Gina Livermore review lessons from SSA's efforts to support youth receiving benefits in transitioning to a successful and more self-sufficient adulthood. These efforts generally estimate mixed effects on benefit receipt, but are promising in improving participants' social connections. Wittenburg and Livermore discuss new policy directions and partnerships SSA could explore. They note that youth needs are different from those of adults and are not separable from families. As a result, services that focus on family outcomes are important. Additionally, interagency collaborations, such as those used in the PROMISE demonstration are important. Given the large number of services and program models for youth, building on existing programs, such as Job Corps, could provide fruitful next steps.

Till von Wachter reviews the importance of looking at the heterogeneous impacts of the demonstrations in Chapter 7. He shows why subgroup impacts are important to estimate. He also provides some suggestions for groups that might be important to look at more closely. von Wachter notes that each demonstration uses different definitions for earnings outcomes, age groups, disability groups, etc., which hinders cross-demonstration comparisons. He notes that standardized outcomes and subgroup definitions would be helpful when comparing across demonstrations. He also describes the state of the art on identifying subgroups with different impacts.

In Chapter 8, Vidya Sundar reviews the use of benefits counseling and case management in SSA's demonstrations. Sundar explores the challenges in measuring the effectiveness of these commonly used but often differently implemented services. She notes that we need more information on the timing and nature of benefits counseling and its interaction with other services (e.g., Vocational Rehabilitation). Additionally, more information is needed on models that focus on sustaining employment rather than just getting a job.

Finally, Michelle Wood and Debra Goetz Engler draw lessons from demonstrations' implementation reports in Chapter 9. Wood and Goetz Engler delve into the qualitative and process analyses to glean lessons from how the beneficiaries and recipients were recruited, how the demonstrations were run, and how the interventions were delivered. They also highlight the importance of recruitment and measurement of fidelity to a program model. On recruitment, they note that dedicated recruitment staff can offer advantages over staff responsible for both recruitment and delivery. They also note that site selection and intervention fidelity are important for structured, specialized services. Wood and Goetz Engler point out tradeoffs between centralized and decentralized funding and implementation with respect to the feasibility of data systems, operational policies, monitoring, and other factors. They note that emergency and basic needs of participants may impede participation and engagement in the intervention. Their insights are helpful for understanding how SSA could implement future demonstration efforts, as well as conduct general outreach.

WHAT'S NEXT

It is our hope that this volume provides a path forward for creating evidence-based policy for SSA's disability programs. There are many important lessons in each of the chapters. Here we provide additional thoughts on four overarching lessons from SSA demonstrations.

1. It would be useful to reset expectations about how many SSDI beneficiaries and SSI recipients will return to work absent a very large program change. The interventions tested so far have not resulted in large numbers of people exiting these programs, even when the programs increased employment. Work can itself be a good outcome, though, even if doesn't constitute sustained SGA in a competitive labor market. It would be helpful to have a fuller conversation about what the goals of interventions should be—program savings or improving the well-being of SSDI beneficiaries and SSI recipients. These two goals are not necessarily mutually exclusive, but there should be discussions about what realistic and meaningful goals a demonstration could attain. It is also important to consider the value in testing the underlying assumptions about SSDI beneficiaries and SSI recipients, program incentives, and work. For example, an "Ultimate Demonstration" (see Gubits et al. 2019) eliminating any effect of earnings increases on benefit receipt (except for as any earnings increase affects medical improvement) could test the assumption that beneficiaries and recipients would return to work in large numbers if all financial disincentives in the program disappeared. Though it may not be a feasible policy to implement nationally, it could provide compelling evidence about how far work incentives simplification could move the needle.
2. Future demonstrations should, when feasible, include additional, meaningful treatment arms to allow SSA to determine the impact of specific intervention

components or alternative policies. Many demonstrations test packages of policies or services, and many of their evaluations have been unable to disentangle the effects of each piece. Understanding which parts of these packages work (and for whom) would enable policymakers to understand whether some parts of interventions could be implemented even when the full intervention could not (or should not). Similarly, knowing about variations of policies (e.g., different offset values or different doses of services) may be more informative than just knowing whether a single package works. Knowing only that a package works or not does not tell policymakers whether it is overprovided, in which case similar effects could be had with fewer resources, or whether the theory is good but the amount of resources necessary makes the package cost prohibitive.

3. New data matches, qualitative data, fidelity metrics, and other information about intervention implementation are often necessary to go beyond the impact analyses presented at the end of a demonstration. How services are delivered is important to understanding why there was or was not an effect. Knowing intervention dosages, how strictly the intervention followed the intended logic models or theories of change, or what the local context to the demonstration was can help policymakers better understand findings. Most SSA demonstrations have included process analyses, but few have included validated measures of fidelity of implementation. It is important to note that such fidelity metrics often require strong up-front planning and clear models that take time to develop and validate. Novel data matches can be pursued even for long-completed demonstrations, to learn more from past investments.
4. Different populations have different needs; targeting can be challenging, but also more effective. Though SSDI and SSI are national programs, some policies or services may be more of an incentive or otherwise effective for certain groups. SSA already acknowledges this in the disability determination process, where those older than age 50 are subject to additional considerations based on their education. Similarly, there are special work incentives for youth and blind individuals. Interventions that focus on specific groups, such as those in the MHTS, SED, YTD, PROMISE, and AB demonstrations might be more productive than policy changes or services that apply to everyone.

One topic not generally covered by the chapters, yet increasingly important, is ensuring diversity in researchers and participants in the demonstrations. Different perspectives bring new ideas to all areas, improving the work, and disability policy is no different. SSA has typically held expert panels prior to conducting demonstrations, and that practice should continue. Ensuring these panels include people with disabilities, people of color, and people with lived experiences in the programs would help improve the value of demonstration designs. Additionally, the teams

implementing and evaluating the demonstrations should be similarly diverse and representative of the populations being studied.

Overall, SSA has shown that it can conduct operational policy demonstrations, service-based demonstrations tailored to local conditions or the national program, nudge-style informational interventions, and a variety of other types of demonstrations. It has partnered with other federal agencies, state and local agencies, community health centers, schools, non-profits, and others. As SSA moves forward in developing new policies, demonstrations can clearly have a role. Used appropriately and judiciously, demonstrations provide rigorous evidence, testing whether the most well-meaning policies have the intended effects and helping to ensure that ineffective or harmful interventions do not become a permanent piece of disability policy.

Chapter 2

Design of Social Security Administration Demonstration Evaluations

Burt S. Barnow
George Washington University
David H. Greenberg
University of Maryland, Baltimore County

An evaluation plan should be developed as the first step in evaluating a program or intervention at the heart of a demonstration. This plan can include decisions about the types of evaluation to conduct (the menu includes *process analysis*, *impact analysis*, and *cost-benefit analysis*). For impact analyses, the plan includes whether to use an experimental design, a quasi-experimental design, or some other approach; how to select the geographic area(s) to include in the evaluation; whom to include in the research population (e.g., everyone affected by the intervention being evaluated or just those who volunteer to participate in the evaluation); the outcomes to assess (e.g., earnings, transfer benefit amounts, health status, mortality); the number of years over which to assess those outcomes; the data to collect or obtain and use (e.g., survey data, administrative data, observation data); and the statistical methods for analysis. (Though we focus on impact evaluations in this chapter, other types of evaluations require similar decisions with analogous considerations.) Decisions concerning these topics can cause enormous variation in how evaluations are conducted and the conclusions that they produce.

The first section of this chapter ("Major Evaluation Design Lessons") discusses these topics, using the evaluation designs from 16 Social Security Administration (SSA) evaluations to illustrate the points we make. These 16 are evaluations for which a published impact evaluation exists and where either the Social Security Disability Insurance (SSDI) program or the Supplemental Security Income (SSI) program was involved. Because the findings from these evaluations are described elsewhere in this book, we do not cover findings here, instead focusing on design and analysis topics. The chapter's second section ("Areas for Further Exploration") discusses some topics about evaluation in practice that so far have garnered little attention in the SSA's evaluations but are worth examining in future evaluations. These topics include alternative experimental designs (e.g., cluster randomization, staggered rollout designs, and factorial designs), rarely estimated effects (e.g., general equilibrium effects, entry effects, program components effects), and site representativeness. The chapter's final section presents our conclusions.

Throughout, we suggest options that we believe might improve the evaluations. These suggestions are not meant as criticisms of past evaluations (evaluation reports

do not always describe all the designs considered but not implemented or the reasons that particular designs were adopted); instead they are used to flag future opportunities.

The 16 evaluations we reviewed are listed in Exhibit 2.1.

Exhibit 2.1. Reviewed SSA Evaluations

Non-Experimental
Proof-of-Concept Studies
Benefits Entitlement Services Team (BEST) demonstration
Homeless with Schizophrenia Presumptive Disability (HSPD) Pilot demonstration
Impact Analyses
Homeless Outreach Projects and Evaluation (HOPE) demonstration
State Partnership Initiatives' SSI Work Incentives Demonstration Project[a]
Experimental
Classical Experiments
Transitional Employment Training Demonstration (TETD)
Project NetWork demonstration
Accelerated Benefits (AB) demonstration
Benefit Offset Pilot Demonstration (BOPD)
Benefit Offset National Demonstration (BOND)
Mental Health Treatment Study (MHTS) demonstration
Youth Transition Demonstration (YTD)
Promoting Readiness of Minors in SSI (PROMISE) demonstration
Promoting Opportunity Demonstration (POD)
Demonstration to Maintain Independence and Employment (DMIE)
Nudging Timely Wage Reporting experiment
Natural Experiment
Ticket to Work program

[a] Implemented by the State Partnership Initiative (SPI) in California, New York, Vermont, and Wisconsin (Kregel 2006a). Also known as the SSI Waiver Demonstration Project.

MAJOR EVALUATION DESIGN LESSONS FROM THE SSA EVALUATIONS

The unit of analysis in the 16 evaluations we review is individuals who were receiving or potentially eligible to receive SSDI, SSI, or both. All but 2 of the 16 estimated the impacts of the demonstration's interventions, although most addressed other questions, as well. Consequently, most of this section focuses on estimating the impacts of the program innovations evaluated in the SSA evaluations. However, near the end of the section, we briefly discuss the roles of process analyses and cost-benefit analyses in the SSA evaluations. Process analyses are essential for interpreting impact estimates, and impact estimates are key ingredients of cost-benefit analyses.

To estimate the impacts of an intervention, an evaluation must make comparisons between a treated and untreated state. The "treated" state is the exposing of individuals (the "treatment group") to the intervention itself or to an offer of it. The "untreated" state is the withholding of the intervention. Evaluators call the untreated state the

"counterfactual" and use it to determine what would have happened in the absence of the intervention.

Of the 14 impact evaluations we reviewed, 12 based their comparisons on an "experimental" design, meaning participants in the evaluation (the "research sample") were randomly assigned in a lottery-like process either to one or more treatment groups or to a "control group" that continued to be subject to the policies or programs that already existed (the counterfactual). In an experimental design, random assignment ensures that the treatment group(s) is initially similar to the control group. As a result, any measured difference in outcomes between the treatment and control groups can be attributed to the intervention: that is, the treatment caused the difference (on average).

The two other impact evaluations relied on "quasi-experimental" designs, which still made comparisons between the treatment and counterfactual conditions, but they did not use random assignment to allocate evaluation participants between the treatment group and a "comparison" group. In the quasi-experiments, evaluators made attempts to adjust for any initial differences between the groups being compared.

Non-Experimental Designs

"Non-experimental designs" refers to evaluations in which there was no randomized control group. Of the 16 SSA evaluations we reviewed, four were non-experimental. Two of these attempted to estimate impacts (as such, they can be classified as "quasi-experimental," as discussed above) and two did not attempt to estimate impacts. These latter two were "proof-of-concept" studies. Because most of this chapter is concerned with impact analysis, we first briefly describe the two non-experimental evaluations that did not attempt to estimate impact and then discuss the two that did in greater detail.

Proof-of-Concept Studies

The Benefits Entitlement Services Team (BEST) demonstration project examined whether homeless SSI and SSDI applicants in Los Angeles County could achieve faster determinations and increased program entry. The Homeless with Schizophrenia Presumptive Disability (HSPD) Pilot evaluation, located in three offices in Northern California, also aimed to achieve faster determinations for homeless SSI applicants, as well as higher payment amounts. BEST had no comparison group; as a result, program impacts could not be estimated, and the evaluation made no causal claims (Kennedy and King 2014). HSPD had three comparison groups, comprising individuals with similar diagnoses as those in the treatment group but who did not receive assistance in the SSI application process. Differences between the treatment group's and comparison group's outcomes were calculated, and *t*-tests were used to gauge statistical significance. However, the evaluation did not attempt to control for underlying differences in characteristics between the groups, and the evaluation report

made no causal claims (Bailey, Goetz Engler, and Hemmeter 2016). The main objective of the HSPD evaluation was to see whether the treatment could be successfully implemented, not to estimate impacts.

Although neither of these evaluations claimed to estimate causal impacts, they both provided other valuable information. Proof-of-concept studies such as these are a useful first step in developing a new program or approach, to see whether it can be successfully implemented. After a program is successfully implemented, an impact study can be considered.

Non-Experimental Impact Studies

We now turn to the two non-experimental evaluations that did estimate impacts. Because the groups are not constructed through random assignment, they likely differ in ways that will affect their outcomes but for reasons not attributable to the treatment. For example, average post-program earnings might differ between the treatment and comparison groups because of differences in their education or motivation. If these differences are not taken into account, the impact estimates will be biased. That is, some of what we call the "impact" will be attributable to the program; some of it will be attributable to the groups' differences in education, motivation, and so on. Consequently, it is essential in estimating impacts in non-experimental evaluations to adjust for differences in the treatment and comparison groups' characteristics.

There are several ways to make such adjustments. We next briefly describe four approaches that are common—use of control variables, propensity score methods, difference-in-differences analysis, and regression discontinuity analysis—and then describe the extent to which the two SSA quasi-experimental evaluations successfully controlled for differences between treatment and comparison groups.

Use of Control Variables

Most evaluations have available various measures of the research sample's characteristics prior to beginning the treatment. Such characteristics might be, for example, their demographics (age, gender, race/ethnicity, etc.), education, and previous work experience. Various statistical techniques, with regression analysis perhaps the most frequently used, can adjust for differences among individuals in these characteristics. This approach has some important limitations. One is that the variables could have been inaccurately measured; even random measurement error of an independent variable can bias estimates of the treatment impact.[1] Second, the way the variables are used to make the adjustment may not be correct. For instance, each year of education prior to the treatment might be assumed to have the same impact on

[1] Random measurement error does not lead to biased estimates of treatment impacts when study participants are assigned to treatment status randomly; but in non-experimental evaluations, the coefficients could be biased. See, for example, Barnow (1976).

earnings, when the 12th year actually has a greater impact than the 11th year. More important, measures of some potentially important variables, such as motivation, might not be available. In the evaluation literature, such internal characteristics are known as "non-observables" (e.g., motivation) as opposed to "observables" (e.g., years of education).

Propensity Score Methods

Propensity score matching involves statistically matching or weighting members of a potential comparison group to individuals in the treatment group on the basis of their observable characteristics. In other words, each member of a treatment group is paired with one or more potential members of a comparison group on the basis of the similarity of those characteristics. The closer the match, the higher the score. Those individuals with the highest scores become members of the comparison group; the remainder of the observations are discarded.[2] Propensity score matching is subject to the same limitations as the use of control variables: measurement errors in the variables used for matching, how these variables are specified, and the unavailability of non-observables. There is evidence that considerable bias sometimes continues to exist even after propensity score matching has been done because some of the differences between the treatment and comparison group can remain (Smith and Todd 2005; Wilde and Hollister 2007). King and Nielsen (2019) suggest methods that can be used to avoid this drawback.

Difference-in-Differences Analysis

If data are available to determine pre-treatment levels of the outcome variables as well as post-treatment outcomes for both the treatment and comparison groups, a difference-in-differences analysis can be performed. This is the analysis approach that is used with a pretest-posttest comparison group design (Shadish, Cook, and Campbell 2002). Although difference-in-differences analysis can be somewhat complex in practice, the basic idea is to net out the pre-treatment differences in outcomes between the treatment and comparison groups from their post-treatment differences in outcomes (Gertler et al. 2011, chap. 6). For example, if the annual post-treatment earnings of the treatment group are $1,000 larger than the annual post-treatment earnings of the comparison group, but the pre-treatment difference between the groups was $300, a simple difference-in-differences estimate would imply that the net impact of the treatment is $700.

Although this approach is quite powerful and is widely used, it will be incorrect to the extent that some factor other than the treatment influences the post-treatment difference between the treatment group and the comparison group (e.g., the treatment

[2] Guidance on using propensity score matching can be found in Caliendo and Kopeinig, (2008).

group lived in a state that raised its minimum wage and the comparison group lived in a state that did not). If such other factors are present, then estimates of the differences between the groups will be biased (Wing, Simon, and Bello-Gomez 2018).

Regression Discontinuity

The regression discontinuity design, which can also be complex in practice, requires that individuals be assigned to the treatment group and comparison group based on their score on some known and non-manipulable measure. For example, individuals were assigned based on a score for the severity of a disability—with those on one side of the cutoff designated to receive the treatment and those on the other side of the cutoff designated to not receive it (see Imbens and Lemieux 2008; Bloom 2009). Individuals near the cutoff are likely to be very similar, allowing those just above and just below it to be appropriately compared. Encouraging evidence exists that regression discontinuity can produce findings that are similar to those resulting from experimental designs (see Cook, Shadish, and Wong 2008). However, regression discontinuity is limited to evaluations in which a score has been used for assignment purposes, which occurs relatively rarely.

Two Examples

Given this background, consider the two non-experimental SSA evaluations that estimated impacts. Both had comparison groups that were very different from the treatment groups, but they did not make use of propensity score matching or difference-in-differences analysis, and they could not use a regression discontinuity design because scores were not used to assign the groups. As discussed in greater detail below, this suggests that the findings from these two evaluations are limited.

The Homeless Outreach Projects and Evaluation (HOPE) treatment was implemented in 41 grantee agencies that assisted individuals with disabilities experiencing homelessness in applying for SSI or SSDI. Like BEST and HSPD, HOPE funded the agencies to attempt to reduce processing time and claim denials. The comparison group was composed of individuals with disabilities experiencing homelessness at 32 similar agencies that did not receive HOPE funding (McCoy et al. 2007). Although the agencies were directly subject to the treatment (receiving HOPE grants), the objective was to improve the situation for their clients. Consequently, in conducting the analysis, the evaluation compared the clients, not the agencies that served them.

In identifying a reasonable comparison group, the evaluators attempted to select comparison agencies that had characteristics similar to those of the treatment agencies (e.g., in location, agency size, and populations served). The evaluation report did not indicate how successful they were in matching the treatment sites along these lines. Moreover, there is still the question of why the treatment agencies had received HOPE funding and the comparison agencies had not. Although the agencies might have been

matched on measurable characteristics, the non-observable characteristics were possibly important and related to outcomes.

The evaluators compared the characteristics of clients at the two sets of agencies, reporting "no [statistically] significant differences" (McCoy et al. 2007, xii). The evaluation used regression analysis to control for differences in individual applicant characteristics between the two groups in estimating impacts on time until benefit determination and claim denials. However, there were some serious data problems. Although the HOPE agencies provided records for 3,055 clients, the comparison agencies provided only 214 records. Beyond the differences in characteristics of agencies and their clients, this major difference in data coverage implies additional potential bias in the impact estimates. In the future, given similar circumstance, SSA might consider using financial incentives in exchange for agencies providing high-quality administrative records.[3]

In additional analysis, the HOPE evaluation made a pre-treatment/post-treatment comparison of the housing situation of clients at the treatment agencies. The problem with this comparison is that the housing situation for at least some individuals who were homeless at the beginning of treatment might be expected to improve even in the absence of treatment, a phenomenon sometimes referred to as "regression to the mean." This impact might have been better estimated with a difference-in-differences approach, in which the pre-treatment difference in housing situation between the treatment and comparison groups was netted out of the post-treatment difference between the two groups. Doing this would have required information on both the pre- and post-treatment housing situation for the comparison group, but the evaluators did not have this housing information. It is not clear whether the comparison agencies collected these data.

An alternative approach to the HOPE evaluation would have been to select treatment and comparison agencies when the program was first initiated in 2004, perhaps by random assignment. Another, perhaps more feasible possibility would have been to have the agencies that wished to adopt HOPE to roll it out randomly, and then compare clients at the early rollouts with those at the late rollouts. This is a type of "stepped-wedge" design, which we further discuss in the next section ("Areas for Further Evaluation"). To use either a random assignment or stepped-wedge approach, the assignment mechanism must be incorporated into the evaluation design prior to program implementation. This was not done in the case of the HOPE evaluation, possibly because the decision to conduct an evaluation was not made until after the treatment was implemented.

SPI's SSI Work Incentives Demonstration Project (also called "SSI Waiver Demonstration Project") implemented four waivers intended to encourage

[3] If all the agencies involved in a demonstration are under contract, then a requirement to provide high-quality data can be written into the contract. However, HOPE was operated under a grant, not a contract. Moreover, the agencies asked to provide data on the comparison group were not part of the grant.

employment among SSI recipients by providing financial incentives to those who volunteered to be subject to the waivers. Incentives included, for example, cutting the SSI benefit reduction rate (BRR) for earned income in half. All four waivers were implemented in three states (California, New York, and Wisconsin), and three of the waivers were implemented in a fourth state (Vermont).

As in the case of HOPE, once the intervention had been implemented, it was too late to use an experimental design, necessitating creation of a comparison group. The evaluator used two alternative comparison groups: (1) SSI recipients in the waiver states who were not subject to the waivers because they did not volunteer to participate in the demonstration; and (2) SSI recipients in eight non-waiver states that, like the four waiver states, received funding under the SPI, but did not implement the waivers.

Because of limited sample size, the data were pooled across the four treatment states and the eight comparison states. Key program impacts that were estimated included employment status and gross earnings. In the analyses involving the two comparison groups, the evaluator controlled for demographic differences between the treatment and comparison group members and for their pre-intake education, training, and employment (Kregel 2006b).

The two comparisons used in the SPI impact study have a number of shortcomings:

- Non-observable differences between the treatment and comparison groups might have affected comparisons between the groups' outcomes. A difference-in-differences approach could have been used to account for non-observable differences. It is not clear why this approach was not used; the needed data did apparently exist. However, perhaps the use of pre-treatment outcomes in the regression equations was sufficient.
- The use of volunteers for the treatment group poses a challenge: the treatment group includes only volunteers, whereas the comparison groups include only non-volunteers of two types. Within states, volunteers were compared to non-volunteers; across states, volunteers were compared to SSI recipients, only some of whom would have been volunteers if they had had the option. Propensity score methods could have been used to improve the match between the treatment and the comparison groups.
- Contextual differences exist between the waiver and non-waiver states; and differences in how they administered the non-waiver components of the SPI, primarily benefits counseling, might have affected the impact estimates. These differences were not taken into account in conducting the impact analysis.
- A comparison group did not exist for New York. Because of this and other problems with estimating impacts for New York, the state could have been dropped from the analysis, or a sensitivity analysis could have been conducted with New York omitted. However, New York accounted for about

half the available treatment group observations, so omitting it would have resulted in dropping a major portion of the treatment group.

- Although all four treatment states were pooled for purposes of analysis, Vermont did not have one of the waivers, whereas the other three states had all four. Moreover, there were differences among the states in how they implemented the waivers.[4]

Given the potentially severe problems listed above, it would have been much better to have used a randomized evaluation design to evaluate the waivers implemented in the four treatment states. For maximum learning, a multi-armed experiment could have been used. However, as discussed next, random assignment (multi-armed or not) is not always feasible. In the case of the SPI project, a decision to use random assignment would have had to be made prior to implementing the intervention but was not.

A key lesson from these two evaluations for future evaluation is this: it is markedly more difficult to adequately evaluate retrospectively than prospectively. Evaluations planned prospectively are much more likely to be able to incorporate random assignment and thereby produce unbiased impact estimates.

Experimental Designs

Unlike non-experimental evaluation designs, randomized (experimental) evaluation designs prompt much less concern about differences unrelated to the treatment occurring between the groups being compared, except by chance alone. Nonetheless, challenges also arise. This subsection first discusses some issues concerning the use of randomized designs and then describes the key features of some of the 12 SSA experimental evaluations as a means for introducing the challenges confronted and lessons suggested by this rich body of past work.

The Pros and Cons of Social Experimentation

There is a substantial literature, and substantial spirited discussion, on the merits of using experimental evaluations for impact analyses. This chapter is not the place to air the full debate, but we raise some of the key issues.

Burtless (1995) argues that experimental evaluations have several strengths. The design:

- ensures the direction of causality;
- ensures the absence of selection bias, which can cause incorrect estimates of impact;

[4] Pooling is further discussed in the subsection "Pooling across Sites."

- permits tests of treatments that do not naturally occur; and
- makes findings persuasive to policymakers and the public.

In part because of these strengths, government clearinghouses, such as the US Department of Labor's Clearinghouse for Labor Evaluation and Research (CLEAR) and the US Department of Education's What Works Clearinghouse, generally provide higher quality ratings to experimental evaluations over other designs, if the evaluations meet other important criteria.[5]

Literature disputing the superiority of experimental evaluations falls in two categories—practical issues and technical issues.[6] Practical arguments against experimental evaluations include these:

- Random assignment in ongoing programs can be disruptive; similar individuals in the same offices must be treated differently.
- Experimental evaluations require more time to arrange for sites to be selected and enrolled and mechanisms installed for implementing random assignment.
- Random assignment in some programs is illegal if the authorizing legislation mandates that everyone eligible must receive the program.
- Random assignment to some programs is unethical.[7]

The technical arguments against random assignment generally contend that the assumptions required for an experimental evaluation to generate unbiased estimates of program impacts are often not met.[8]

It is our contention that, when legal and ethical, experiments can overcome their shortcomings and provide strong evidence for policy decisions. We discuss the experimental design and its merits because SSA has done an admirable job over the past nearly four decades using experimental evaluations as a means to uncover the impacts of potential policy changes. The consistent use of experimental evaluations has provided a strong evidence base for assessing alternative program strategies. Our recommendation is that SSA continue to prioritize use of experimental evaluation

---

5 For example, the criteria for a high rating in CLEAR is as follows: "A high rating means we are confident that the estimated effects are solely attributable to the intervention examined. Two types of studies can receive a high rating: (1) well-conducted [randomized control trials] that have low attrition and no other threats to study validity and (2) [interrupted time series] designs with sufficient replication wherein the intervention condition is intentionally manipulated by the researcher. [Such] designs that do not qualify for a high rating can be evaluated against CLEAR's evidence guidelines for regression analyses" (DOL 2015).

6 Bell and Peck (2016a) suggest three categories of concerns with experiments (with a total of 15 concerns): ethical, scientific, and feasibility.

7 See, for example, Blustein (2005) for arguments that denying eligibility to participate in the Job Corps in order to conduct an evaluation is unethical.

8 Recent advocates of this position are Deaton and Cartwright (2018) and Cook (2018). The former state their conclusion strongly: "We argue that any special status for [randomized control trials] is unwarranted" (2).

designs; later, in the section "Areas for Further Exploration," we suggest how the agency might push the envelope further.

Examples of SSA Experimental Evaluations

All but one of the 11 SSA evaluations designed as experiments used a simple procedure to assign individuals to treatment groups and control groups. The random assignment procedure is essentially a toss of a fair die that makes the pre-treatment characteristics between the groups, whether characteristics are observed or not, the same on average. As a result, any differences in post-treatment behavior between the groups can be attributed to the treatment, rather than to preexisting differences. (In the "Areas for Further Exploration" section of the chapter, we discuss some alternatives to the simple random assignment design.)

To highlight how the experimental evaluation design works in practice, we briefly introduce six of the SSA experiments in the remainder of this subsection, highlighting their unique features to lend insight into some of the creative things evaluators can do. The following subsections discuss many of the challenges these experiments confront.

Ticket to Work. The Ticket to Work program provided SSI recipients and SSDI beneficiaries "tickets" that they could give to vendors in exchange for providing them with services and training to assist them in obtaining employment. The evaluation was of an actual program that was just being rolled out. For that reason, instead of being based on the simple experimental design just described, the evaluation exploited that the timing of when SSDI beneficiaries and SSI recipients received their ticket was essentially random. This was because, as SSA has done in several projects, the queue for receiving a ticket was determined by the last digit of a beneficiary's or recipient's Social Security number (which is essentially random). Outcomes for those who received their ticket earlier were compared to outcomes for those who received their ticket later (Livermore et al. 2013). Thus, there was not a control group in the usual sense. The evaluation of Ticket to Work is interesting because instead of purposefully randomly assigning individuals to treatment and control groups for evaluation purposes, it took advantage of a program feature that existed for other reasons.[9] This is sometimes called a "natural experiment."

Project NetWork. Project NetWork, which experimentally tested case management as a means of promoting employment among SSI recipients and SSDI beneficiaries, had an unusual non-experimental design feature: four different models for providing services were tested, with each tested in two of eight sites (Kornfeld et al. 1999). However, because only a single treatment was tested in each site, differences in how the intervention performed could be assessed only by non-experimental inter-

[9] One can argue that, technically, Ticket to Work is not an experiment because group assignment is not random; however, because group assignment is based on the final digit in the Social Security number, which is assigned randomly, we are treating the program as an experimental evaluation design here.

site comparisons. As a result, any inter-site differences in impacts might be attributable to site differences in the characteristics of the participants or in the economic environment, rather than differences in the tested intervention. More sites per model might have improved these comparisons, but this would have increased the cost of the evaluation and might not have been feasible for budgetary reasons.[10] A multi-armed approach, which is described in the following paragraph, could also have been used.

Accelerated Benefits (AB). A major evaluation design difference among the SSA demonstrations is the number of interventions tested in each evaluation site. Although most evaluations had only a single treatment arm, three evaluations had two arms, and one had four arms. Outcomes for these additional treatment groups could be compared not only to outcomes for a control group but also to one another. For example, the Accelerated Benefits demonstration was fielded to address the fact that SSDI beneficiaries had a two-year waiting period before they could qualify for Medicare. The demonstration had two treatment arms: AB and AB Plus. SSDI beneficiaries were randomly assigned among the two treatment arms and a control group. Both treatment arms provided health benefits to SSDI beneficiaries who were in the waiting period and were otherwise uninsured. Those beneficiaries randomly assigned to the AB Plus treatment arm additionally qualified for certain services provided by telephone, such as employment counseling (Michalopoulos et al. 2011). By comparing outcomes (e.g., earnings, SSDI payment amounts) for the two treatment groups, it was possible to determine whether availability of the additional telephone services had impacts over and above impacts resulting from the provided health benefits.

Benefit Offset National Demonstration (BOND). BOND is one of several SSA demonstrations that tested the impacts of replacing the SSDI cash cliff (an earnings threshold at which benefits become zero) with a 50 percent BRR. BOND involved two parallel experiments: Stage 1 targeted the entire SSDI population within the study sites, whereas Stage 2 targeted only volunteers. Stage 2 of BOND also had two treatment arms: one group received enhanced work incentives counseling, whereas the other group received standard work incentives counseling. By comparing these two groups, the evaluation could determine any added impact of enhanced counseling (Gubits et al. 2018a/b).

Promoting Opportunity Demonstration (POD). Like BOND, the currently running POD is testing replacing the threshold at which all SSDI benefits cease with a 50 percent BRR. However, the POD threshold is lower than the BOND threshold. In addition, it also is testing eliminating the nine-month Trial Work Period (TWP) and the three-month Grace Period under the existing SSDI program, during which beneficiaries are not subject to a BRR. Also, like Stage 2 of BOND, POD has two treatment arms. SSDI benefits are suspended for individuals randomly assigned to the first arm if their earnings are sufficiently large that their benefits reach $0 (called the

[10] With a sufficiently large number of sites, it could be possible to pool across the sites and tease out the separate impacts of the various program features. For example, see Bloom, Hill, and Riccio (2003); Greenberg, Meyer, and Wiseman (1993, 1994).

"full-offset point"). They can, however, again receive SSDI if their earnings subsequently fall below the full-offset point, without having to re-enroll in the program. Beneficiaries randomly assigned to the second arm have their SSDI entitlement terminated when their earnings reach the full-offset point for 12 consecutive months. As a consequence, they need to reapply for SSDI if their earnings subsequently fall below the full-offset point for 12 consecutive months, although they are eligible for expedited reinstatement of benefits (Hock, Wittenburg, and Levere 2020). Thus, the second treatment could reduce the SSDI rolls by a greater amount than the first treatment.[11]

Nudging Timely Wage Reporting. This experiment, run by SSA staff and academic researchers associated with the White House's Social and Behavioral Sciences Team, involved sending a letter to SSI recipients reminding them of their wage reporting responsibilities. The evaluation involved a control group plus four treatment arms, with the letter's language varying among the arms: (1) simple information about reporting (included in all letters); (2) social information on reporting behavior; (3) information increasing the saliency of the penalties for non-compliance; or (4) both social information and information on penalties (Zhang et al. 2020). With this design, it was possible to determine whether the specific content of the letter made a difference. A nudge experiment such as this can provide considerable information inexpensively and should be encouraged.[12] Unlike most of the SSI evaluations, participation in the treatment groups was not voluntary, as in Stage 1 of BOND.

Sample Design Issues: Statistical Power and Minimum Detectable Effects

SSA's prior demonstrations had a large range in sample size. The largest studies were Stage 1 of BOND, which had a treatment group of 77,101 and a control group of 891,429, and the Nudging Timely Wage Reporting experiment, which included 50,000 participants in four treatment groups and a control group. At the other extreme, the Centers for Medicare and Medicaid Services (CMS)–sponsored Demonstration to Maintain Independence and Employment (DMIE) had 184 participants in one state and 500 in another, evenly divided into treatment and control groups.

Assessing whether an evaluation has an adequate sample size to permit detection of policy-relevant impacts is complex. It depends on a number of parameters including tolerance for Type I and Type II errors,[13] whether the evaluation uses an experimental

---

[11] For further detail about the work incentives features of the existing SSDI program and how POD modifies them, see the *Red Book* (SSA 2020e) at https://www.ssa.gov/redbook/.

[12] SSA implemented three other "nudge" experiments that involved varying the language in notices sent to beneficiaries. On the US General Services Administration/Office of Evaluation Sciences website (https://oes.gsa.gov/), see "Increasing SSI Uptake among a Potentially Eligible Population"; "Increasing Participation in Ticket to Work"; and "Communicating Employment Supports to Denied Disability Insurance Applicants."

[13] A Type I error is rejecting the null hypothesis of no effect when it is true, and a Type II error is failing to reject the null hypothesis when it is false.

design, the allocation of the sample between treatment and control status, and the actual program impact.[14] Bloom (1995) developed a framework for analyzing statistical power issues so that evaluators can calculate the minimum detectable effect and/or the minimum required sample size.[15] Bloom frames the analysis as follows:

> The minimum detectable effect of an experiment is the smallest effect that, if true, has an X% chance of producing an impact estimate that is statistically significant at the Y level. X is the statistical power of the experiment for an alternative hypothesis equal to the minimum detectable effect. Y is the level of statistical significance used to decide whether or not a true effect exists. (547)

Bloom's equations inform the sample size that would produce a given statistically significant impact and the impact that would be detectable for a certain sample size.

Most of the SSA evaluations reported that a power analysis was performed as part of their planning; examples include the Youth Transition Demonstration (YTD) and Promoting Readiness of Minors in SSI (PROMISE). Most of the evaluations had a large enough sample that if the intervention being evaluated achieved the anticipated impact, the results would be detected as statistically significant. However, a few demonstrations had too small a sample to be expected to detect statistically significant findings if the intervention was as effective as anticipated. The reasons for inadequate sample sizes are predictable, and the most common was insufficient resources. For example, Michalopoulos et al. (2011) did a power analysis and determined that the AB demonstration needed a sample of 2,000 participants, but one of the treatment arms cost more than anticipated. Consequently, the allocation of the sample was modified, and much of the analysis used a sample of only 1,531 participants.

The DMIE also had relatively small samples in participating states (Whalen et al. 2012). In DMIE, four states developed strategies to assist individuals with specified disabilities to remain off SSI and SSDI. The selected disabilities varied across the states, which made pooling across states of questionable value. Hawaii targeted people with diabetes; Kansas, individuals with a variety of physical and mental conditions; and Minnesota and Texas, people with behavioral health issues. Although Minnesota and Texas had more than 1,000 participants in their treatment and control groups, Kansas had 500, and Hawaii had only 184. The evaluation report notes that the sample sizes might not be adequate to achieve statistically significant findings of the magnitude expected for the results to be policy relevant, but there is no discussion of whether a power analysis was conducted beforehand. In some of the DMIE states, the

---

[14] In evaluations in which the objective is to determine whether a program can be successfully implemented, rather than to estimate the program's impact, the desired sample size is not determined by statistical criteria.

[15] In addition to Bloom (1995), the concepts are explained, for example, by Dong and Maynard (2013) and Orr (1999).

sample was large enough for an overall impact analysis, but not large enough to conduct subgroup analyses, which might offer policy-relevant results.

Project NetWork had an overall sample of 8,248 individuals randomly assigned to treatment and control groups (Kornfeld and Rupp 2000). The demonstration tested four delivery models in two states each, and the participants had a wide range of disabilities. Kornfeld and Rupp warn: "Interpreting estimated impacts for subgroups requires caution. Whenever we analyze impacts for subgroups, the sample size declines, and the standard errors of estimates for many of the subgroups become quite large, so that only large impacts could be detected as statistically significant" (24).

Population-Representativeness

To produce impact estimates that are valid for an entire target population, an evaluation needs to include as representative a sample of that target population as possible. In the words of Stapleton et al. (2020), the sample used in an evaluation needs to be "population-representative."

There are several reasons the sample used in an evaluation might not be population-representative. Two of these are discussed below. The first is that the research sample could be located in sites that are not representative of the population of potential program participants nationwide. The second reason is applicable to evaluations of demonstration programs when participation in the demonstration is voluntary. In such circumstances, there is often interest in using findings from the evaluation to predict what would happen if the demonstration program were rolled out nationally and participation became mandatory. As discussed below, it is difficult to extrapolate from findings that pertain to a voluntary program to one that is mandatory. (Of course, the long-run impacts of national programs may also be missed in evaluations of demonstration programs because they operate at a larger scale, information feedback occurs over time, changes in the economy occur, and numerous other considerations. We are abstracting from these considerations in this discussion.)

The Ticket to Work evaluation and the Nudging Timely Wage Reporting experiment were both national in scope. So, too, was Stage 1 of BOND in that it randomly selected its evaluation sites to be nationally representative. Consequently, the samples used in these three evaluations were geographically representative of the national population of SSDI beneficiaries and SSI recipients. This is important because very few evaluations of social programs are based on nationally representative samples.

With the exception of those three, all the SSA evaluations we examined provided services or financial incentives that could be received only by individuals who first

volunteered.[16] If the point of an evaluation is to estimate impacts that are predictive of an ongoing, national program, then evaluations that use volunteers cannot be population-representative unless the ongoing national program would also be voluntary. For example, three of the non-experimental evaluations we reviewed (BEST, HOPE, and HSPD) examined whether the SSI or SSDI application process could be improved for individuals with disabilities experiencing homelessness. Findings concerning this application process can be applicable only to persons who experience homelessness and who volunteer to participate in the demonstration. The DMIE was evaluated experimentally, but like BEST, HOPE, and HSPD, it served individuals who had not yet applied for disability benefits, making mandatory participation infeasible (Whalen et al. 2012). If the volunteers for the demonstration were representative of those who would volunteer in a national program, then the research sample was population-representative.

Some of the other voluntary programs that we reviewed provide an important distinction, as they were demonstrations for which participation could be made mandatory if the services they offered were rolled out nationally (although there is no way of knowing whether this would actually occur). Because these programs had low take-up rates, their research samples would be unlikely to be population-representative

[16] By law, only volunteers can participate in SSA demonstrations that require waiver authority. This was not the case when BOND was implemented, and Stage 1 of BOND is mandatory. Ticket to Work is an evaluation of an ongoing program, not a demonstration; as a result, it was not limited to volunteers. The Nudging Timely Wage Reporting experiment, as well as the three other nudging experiments that are mentioned in note 12, did not require volunteers because they did not require waiving program rules.

Interestingly, although the Nudging Timely Wage Reporting experiment was mandatory and national in scope, it was not population-representative. The sample for inclusion in the evaluation was based on a score, which was developed by SSA to select individuals for a redetermination of their SSI benefits on the basis of the likelihood that their benefits would change. Individuals with the highest scores were excluded from the experiment because they would all be called for a redetermination. The 50,000 individuals in the group with the next-highest scores were included in the experiment, and they were randomly assigned to receive one of the four types of letters or no letter (Zhang et al. 2020). Unfortunately, the results cannot be generalized to SSI recipients with lower scores. An alternative strategy might have been to stratify the total eligible population and then randomly select individuals from each stratum, perhaps assigning those with higher scores a higher probability of being selected. This type of design would permit the evaluators to determine whether the benefit of the intervention varied by score, allowing a policy to be implemented that focused on those most likely to be affected. It was not possible to do this, however, because the full sample was not available to the evaluators.

of mandatory versions of the same programs.[17] The low take-up rates in evaluations of these voluntary demonstration programs provide important information for policymakers considering rolling out the programs nationally, as long as the national version would continue to be limited to those seeking the services the programs provide. The YTD provided waivers of certain program rules that were intended to encourage work. These rules could potentially become part of a national program. Moreover, the voluntary enrollees in the YTD tended to be especially motivated to work. This could have resulted in impact estimates different than what would have occurred had the research sample been more representative of the general population of youth with disabilities who would have been covered by the waivers.

The AB demonstration provided health benefits for new SSDI beneficiaries who did not have private health insurance and were subject to a waiting period before they could qualify for Medicare; such benefits could potentially be rolled out nationally. Fortunately, the treatment was so generous that nearly everyone eligible enrolled. However, 87 percent of those who would have been eligible for benefits under AB already had health insurance; as a consequence, they were ineligible to enroll. In a national program, some new enrollees might leave their existing health plans prior to becoming a beneficiary if they can obtain benefits that are more generous (and the tested plan was relatively generous). Consequently, if tested again, SSA might consider allowing a random subset of individuals who already had health insurance prior to treatment to enroll in the test to see how many will substitute the program's health plan for their own.

Policies that provide incentives to work by changing the SSDI BRR have also been evaluated experimentally by recruiting volunteers. The POD, which under existing law must be evaluated with volunteers, is an important example. In addition, only those who volunteered to participate in the State Partnership Initiative demonstration were eligible for work incentives waivers provided by the SPI project. If rolled out nationally, these evaluated policies could well be available to all SSDI beneficiaries and SSI recipients who meet certain eligibility criteria, not only to those

[17] For example, the Transitional Employment Training Demonstration, which in 1985 was targeted at what was then termed "mentally retarded" SSI recipients, enrolled only about 5 percent of those eligible (Thornton and Decker 1989). Project NetWork is another example of a voluntary program with a low take-up rate: only 5.6 percent of the eligible SSI recipients and SSDI beneficiaries volunteered. It is not surprising that individuals who have a disability that makes working difficult rarely volunteer for programs intended to get them off the SSDI or SSI rolls, especially because they would lose health insurance and guaranteed income (Kornfeld et al. 1999). The YTD enrolled 16-30 percent of eligible youth at its six sites after the evaluators worked "very hard" to attract volunteers (Fraker et al. 2014, xxiii). The evaluation of the MHTS, which attempted to increase employment among SSDI beneficiaries with schizophrenia or an affective disorder, concluded that were it voluntary, "SSA could expect 14 percent of the SSDI beneficiaries with schizophrenia or an affective disorder might enroll in an MHTS-like program" (Frey et al. 2011, 9-5). Not enrolling in MHTS was often due to health constraints and general lack of interest.

who would volunteer for a demonstration. If so, the sample populations used in the evaluations might not be population-representative. On the one hand, under a national program, the volunteers would be more likely to work and have their benefits affected by the intervention than those who did not volunteer.[18] On the other hand, some non-volunteers, if subject to a national program, would be affected by the financial incentives and counseling.

Stapleton et al. (2020, 557) point out that an important rationale for evaluations based on volunteers is that they are less expensive to conduct because the evaluations will generally "require a smaller sample size than a population-representative experiment in order to detect an impact for the treatment subjects of any given size, provided that the volunteers attracted to the experiment contain a disproportionately large share of those volunteers for whom the treatment is salient." In the case of one of the outcomes investigated in BOND, for instance, Stapleton et al. (2020) find that a population-representative evaluation would require three times the sample size as a would a volunteer evaluation to obtain the same minimum detectable effect. A larger sample requirement results in both larger implementation costs and larger survey costs. As Stapleton et al. also recognize, however, cost savings from a voluntary evaluation could come at the cost of learning less about what is relevant.

The voluntary nature of POD creates some special problems in providing lessons for a mandatory program. To some extent, POD is a replication of Stage 2 of BOND, with the main differences being a reduction in the earnings threshold at which the BRR becomes operative and the elimination of the TWP and the Grace Period, which existed in BOND and continue to exist in the regular SSDI program. However, during months that beneficiaries would have been using the TWP and the Grace Period under current law, they are worse off under POD. As a result, such beneficiaries are likely to withdraw from POD or not volunteer in the first place. As a consequence, the information that POD can provide about the impacts of eliminating the TWP and the Grace Period for non-volunteers is limited. Under a mandatory national version of POD, some working beneficiaries will still be in their first year of earnings. Their characteristics are likely to differ from characteristics of those who volunteered for the demonstration.

Although it would be useful to randomly test a mandatory version of POD, this cannot be done at present because "SSA's statutory demonstration authority requires

[18] Differences between those who volunteer for a program and those who do not also suggest the dangers in estimating impacts by comparing outcomes for those two groups, as was done in the SPI evaluation. The two groups are not comparable in ways that are difficult to adjust for statistically.

the use of informed volunteers" (Stapleton et al. 2020, 560).[19] The rationale for this provision is the ethical concern that some beneficiaries would be made worse off, which is exactly what the elimination of the TWP and the Grace Period would do under a mandatory POD. However, Stapleton et al. suggest that in considering a policy that might be adopted nationally and thereby affect non-volunteers, "it is arguably more ethical to instead conduct a population-representative [experiment] that does measure the potential harm" (559). Another possibility in testing POD experimentally would have been to have had a second arm of the experiment that does not eliminate the rules that provide the TWP and Grace Period but is voluntary. However, a similar program design was previously tested in Stage 2 of BOND, so a second test might not have been useful.

Outcome Measures

Because the majority of the evaluated programs were intended to help SSDI beneficiaries and SSI recipients do better in the labor market, it is not surprising that the most common outcome measures in the evaluations were employment and earnings. Employment was most commonly measured as a dichotomous variable (i.e., employed or not employed over a calendar quarter or year). In some evaluations, however, employment was measured as the number of hours worked over the period.

Earnings were measured in several ways, most commonly as quarterly or annual earnings.[20] Social Security disability programs (SSDI and SSI) have earnings thresholds that measure whether an applicant's earning capacity is sufficient that they do not qualify for disability benefits. Specifically, "to be eligible for disability benefits, a person must be unable to engage in substantial gainful activity (SGA). A person who is earning more than a certain monthly amount (net of impairment-related work expenses) is ordinarily considered to be engaging in SGA."[21] Evaluation of BOND's predecessor, the Benefit Offset Pilot Demonstration (BOPD), used earnings above the SGA level as well as total earnings as outcome measures. The BOND evaluation used earnings above the SGA level and several other measures that focused on higher earnings. Defining the BOND Yearly Amount as annualized SGA, BOND used the

[19] However, as indicated by the following statement, SSA (2019b) recognizes the limitations of this provision, and it is requesting modification of it under limited circumstances: "We are also limited in our ability to assess how program changes might affect people beyond the subset of the population who volunteered. As a result, the impacts are not easily generalizable to the national population and may not provide the adequate understanding required to make informed decisions about broader policy changes. In the FY 2020 President's Budget, we included a proposal to expand our authorities to allow us, in limited circumstances, to conduct demonstrations with mandatory participation."

[20] This section of the chapter deals with the outcome variables; a later section discusses the use of administrative data versus survey data.

[21] In 2021, SGA for blind applicants is $2,190 per month and $1,310 for applicants who are not blind. Retrieved December 11, 2020. https://www.ssa.gov/oact/cola/sga.html.

percentage of individuals earning two and three times the amount as additional outcome measures.

Some evaluations included benefits paid as an outcome measure. Interpretation of impacts on benefits paid is less straightforward than interpreting impacts on earnings because there are alternative mechanisms by which the intervention can affect benefits; for example, benefits could decrease because of increased work or failure to comply with program rules. Typically, the amount of benefits paid was the outcome variable, but in one case, the Transitional Employment Training Demonstration (TETD), the outcome was receipt of SSI benefits. Evaluations examining benefits paid included AB, BOND, BOPD, DMIE, POD, PROMISE, YTD, and Project NetWork.

Some of the interventions were intended to improve the health of participants, and evaluations of these efforts included measures of participant health as an outcome. For example, AB and DMIE used scores on the SF-12 questionnaire for mental health and physical health as outcomes,[22] and the evaluation of DMIE also used the percentage of participants with limitations in activities of daily living and instrumental activities of daily living as outcomes. The Mental Health Treatment Study (MHTS) and BOND evaluations used the SF-12 to measure physical and mental health; the MHTS also included a quality of life measure as an outcome. The AB demonstration provided health-related benefits to SSDI beneficiaries during the two years they were required to wait to receive Medicare. Health outcome measures in the AB evaluation included unmet medical needs, self-reported health status, and died since random assignment (Michalopoulos et al. 2011, ES-5).

Because people with some disabilities may experience higher mortality if they do not receive the health care and income provided by SSDI and SSI, some evaluations included mortality as an outcome of interest. Examples include HSPD and AB.

Project NetWork used somewhat different measures of health outcomes, but the evaluation notes that the use of self-reported responses "could mean different things to different respondents" (Kornfeld and Rupp 2000, 23). Measures included self-reported health as excellent or very good, self-reported improvement in health since random assignment, having three or more life skills limitations, having three or more functional limitations, the Mini Mental State Evaluation, and the Mental Health Inventory.

Health is clearly a more complex phenomenon than income to measure, and measurement of health status can be expensive if clinical assessments, rather than self-assessments, are used. SSA might want to consider whether sufficient evidence is

[22] The 12-item Short Form Health Survey (SF-12) is a self-reported measure of physical and mental health. Frey et al. (2011, 2-20) state that the SF-12 is not as detailed as the longer SF-36, but it captures eight aspects of physical and mental health: (1) limitations in physical activities due to a health problem; (2) limitations in social activities due to a health problem; (3) limitations in usual role activities due to a physical health problem; (4) limitations in usual role activities due to an emotional problem; (5) pain; (6) general mental health; (7) vitality; and (8) general health perceptions.

available to establish standardized measures of mental and physical health or to confer this status on existing measures.

Some of the evaluated demonstrations tested interventions intended to speed up the application process for SSI and SSDI. These evaluations often focused on the speed of eligibility determination or the approval rate of applications or both. Examples include BEST, HOPE, and HSPD. Of them, BEST used processing time as an outcome measure, HOPE used time until determination, and HSPD used time until adjudication. These are all appropriate outcomes to examine, but the evaluations appear to presume that faster is always better. In future evaluations, SSA might also use measures of decision accuracy.

Impact Estimation Issues

The SSA evaluations we reviewed varied in how they estimated program impacts—for example, in the data and the statistical approach they used, how missing data were treated, whether they pooled across sites in reporting impacts, length of the follow-up period, and determining the statistical significance of impacts when multiple outcomes are examined. To some extent, the variation across evaluations stemmed from both the nature of the interventions and the objectives of the evaluations. These estimation issues are discussed below.

Data Sources

Evaluation designs are shaped by the data available for analysis. An integral component of an evaluation plan involves determining the relevant data that are available, selecting the most appropriate data, and obtaining access to these data. Chapter 3 in this volume discusses how the data available for SSA evaluations can be improved.

Most of the SSA evaluations we reviewed depend heavily on SSA-provided administrative data that evaluators transformed into analysis-ready files. Frequently used examples of these SSA files include the Supplemental Security Record, which provides demographic information, addresses, and benefit payments amounts for SSI recipients; the Master Earnings File, reflecting that earnings and employment are often key outcome variables; the Master Beneficiary Record, which contains benefit information about each claimant who has applied for retirement, survivors, or disability benefits; and the Disability Analysis File, a collection of data records for both SSDI beneficiaries and SSI recipients from various sources. Administrative data from government agencies other than SSA were also used in a few evaluations. For example, the evaluation of the SPI project used Unemployment Insurance (UI) data and state SSI administrative data, using SSI administrative data for only one site (New York); and BEST made use of the Veterans Benefits Administration database.

Most, but not all, the evaluations also collected survey data,[23] typically at the point when participants were enrolled in the evaluation ("at baseline") and then periodically after enrollment. The MHTS is unique because its impact estimates rely almost exclusively on survey data rather than administrative data, although employment and earnings were among the outcomes examined and, as discussed below, SSA administrative data could provide superior measures of these outcomes. MHTS was also notable in how it conducted its surveys. Over a 24-month follow-up period, nine computer-assisted quarterly surveys were conducted, with the interviewers physically located at each site. Though costly, this approach should reduce recall errors and, in principle, improve survey response rates, although at 82 percent for the treatment group and 86 percent for the control group (Frey et al. 2011), the rates were not exceptionally high.

It is generally more costly to conduct surveys in non-voluntary evaluations (i.e., those in which participation in the evaluated programs is mandatory) than in evaluations where participation in the intervention is voluntary. In non-voluntary evaluations, a smaller portion of the treatment group is likely to respond to the offer of the intervention; as a consequence, a larger sample size is needed. Although it is possible to save on survey costs in non-voluntary evaluations by subsampling from among the evaluation participants, doing so can result in imprecise impact estimates, as in fact occurred in Stage 1 of BOND (Stapleton et al. 2020). Moreover, when the intervention is voluntary but the evaluation is mandatory, such as Stage 2 of BOND, contact with members of the sample occurs at enrollment, whereas there may be little contact with many members of a sample in a population-representative evaluation. Moreover, volunteers have already exhibited an interest in the intervention. As a consequence, response rates might be higher in voluntary evaluations than in non-voluntary evaluations. For example, the response rate in the Stage 1 36-month survey was 57 percent, as compared to 84 percent in the corresponding Stage 2 survey (Stapleton et al. 2020).

Unlike administrative data, it is possible to tailor survey data to the specific needs of an evaluation. Survey data were essential to many of the SSA evaluations because they allowed analysis of outcomes that were not available in administrative data. For instance, surveys can collect data on income from sources other than earnings (e.g., child support, self-employment), hours worked, hourly wage rates, motivation, quality of life, health status, the receipt of program services, and the understanding of program rules. To illustrate, using information collected in a survey, MHTS constructed an index to measure program impacts on the self-determination of its target population. However, to keep the survey short, only a limited number of questions could be

[23] Both BEST and HSPD, which were non-experimental, made use of SSA administrative data, but did not collect survey data. Nor did the Nudging experiment that aimed at increasing wage reporting among SSI recipients. It used the Supplemental Security Record to determine whom to target, to obtain the mailing addresses needed to send nudge letters to those targeted, and to determine whether reported earnings increased as a result of the letters.

administered, which "may have resulted in [the index] being less sensitive to the effects of the interventions" (Frey et al. 2011, 145). The self-determination measure used in the evaluation of the PROMISE demonstration was also less useful than anticipated.

Although surveys are essential for collecting information not available in administrative data, they also suffer important disadvantages. Administrative data, already available for non-evaluation purposes, are much less costly than survey data. Because of these lower costs, administrative data are often available at more frequent intervals and they can be used for longer follow-up. For example, SSA researchers extended the original one-year follow-up period for the AB evaluation to three years (Bailey and Weathers 2014), and there are further plans to extend follow-up to over a decade.[24] Similarly, the final report for the YTD had a three-year follow-up period, which was later extended to between five and seven years (depending on the outcome measure), and plans are to extend it further.[25]

Surveys are subject to nonresponse because members of the research sample cannot be found, or they refuse to be interviewed. These nonresponses typically increase over the follow-up period. If nonresponse correlates with treatment assignment, then the resulting impact estimates can be biased. Surveys also tend to be subject to recall error, as well as to simple misreporting. Moreover, there is evidence that some survey respondents report implausibly high hours and earnings, especially pertaining to overtime work (see Barnow and Greenberg 2015). On the other hand, some respondents can fail to recall brief informal jobs or to correctly remember their hours and earnings in occupations that tend to irregular hours. They also can tend to understate transfer payments (Hotz and Scholz 2001), either intentionally or inadvertently.

As summarized by Barnow and Greenberg (2015) considerable research suggests that, on balance, earnings tend to be overreported in surveys by low-income respondents and underreported by higher-income respondents. When this occurs, impacts on earnings in programs targeted at low-income respondents that are estimated by survey data tend to be biased upward, especially if overreporting is larger for treatment groups than for the control/comparison group (Barnow and Greenberg 2015, 2019). This might occur if members of treatment groups are motivated to exaggerate their success in a program, possibly to impress their interviewer (Barnow and Greenberg 2015).

In contrast to the findings summarized by Barnow and Greenberg (2015), a recent comparison of SSA's National Beneficiary Survey with administrative earnings records from its Master Earnings File found that estimated employment rates and earnings levels for SSDI beneficiaries and SSI recipients were consistently higher in administrative data than in survey data (Wittenburg et al. 2018). One possible partial

[24] Robert Weathers II, email with the authors, November 2, 2020.

[25] Jeffrey Hemmeter, email with the authors, November 13, 2020.

explanation for these findings could be that sometimes multiple earners use the same Social Security number, resulting in erroneously high earnings for one person. This would bias impacts on earning estimated with Social Security data upward. Wittenburg et al. speculate that probably a more important factor is recall error among the survey respondents, causing them to miss some of their earnings and jobs in their responses. This appears plausible because, when they do work, SSDI beneficiaries and SSI recipients with disabilities are likely to work part-time or infrequently. This would bias impact estimates made with survey data downward.

Many evaluations of government training programs and welfare-to-work programs have relied on data used in administrating state UI systems. The problem with UI data is that they miss workers who live or work in states other than the one where the evaluated program is located, who are self-employed, or who work in industries not covered by UI. Workers and their earnings are also missed because of errors in their Social Security numbers. The SSA administrative data that are used in most of the evaluations covered in this chapter suffer much less from these common UI data shortcomings because they are national in scope. Moreover, SSA verifies reported Social Security numbers, and SSA administrative data cover more industries than the UI data do.[26] That said, both UI and SSA administrative data miss some government employees and workers paid outside the formal economy. Surveys can capture employment that is not covered in administrative data. Of course, earnings obtained in the informal economy are also unlikely to affect SSDI and SSI benefit levels, complicating how they should be treated in evaluations of SSA programs.

Missing data on workers are important in estimating impacts on employment and earnings with administrative data because when workers do not show up as employed, they are usually treated as nonworkers, thereby biasing the estimates downward (see Barnow and Greenberg 2015, 2019). Such biases are much more important if more workers are missed in the treatment group than in the control/comparison group. This might be the case, for example, if the intervention causes treatment group members to become self-employed (see Barnow and Greenberg 2015, 2019). As suggested above, these biases are likely to occur less often in using SSA administrative data than in using UI data. For example, an experimental evaluation of the Job Corps used data from both sources to estimate program earnings impacts and found larger impacts with the SSA data than with the UI data. After an investigation, the evaluators attributed part of this difference to erroneous Social Security numbers being more likely in the UI data than in the SSA data (Schochet, McConnell, and Burghardt 2003).

As suggested above, survey data can result in earnings impacts that are upward biased, whereas administrative data can result in earnings impacts that are downward biased, although as indicated by the findings of Wittenburg et al. (2018), this is not necessarily the case. In the PROMISE evaluation—the one SSA evaluation that

[26] A limitation of SSA data for research purposes is that there are delays in obtaining earnings data, which are based on tax years and so are annual and not reported until March the following year at the earliest and not considered "complete" until the following February.

estimated earnings impact with both survey and SSA administrative data—impacts on the annual earnings of the youth who were targeted by the intervention were more than twice as large at four of the six evaluation sites when estimated with survey data instead of with administrative data. Impacts at the remaining two sites were very small regardless of the data with which they were estimated (Mamun et al. 2019).

Statistical Approaches to Impact Estimation

Many of SSA's evaluations have used random assignment to assign individuals to treatment or control status, and most of these evaluations used standard statistical approaches.[27] For continuous outcomes, evaluations most commonly used ordinary least squares; for dichotomous outcomes, logistic regression was most common.[28] All the experiments used an intent-to-treat (ITT) approach, in which the analysis was based on the treatment assigned regardless of whether the treatment group member took up the offer of treatment. In addition, evaluations can compute the average treatment-on-the-treated effect (TOT), in which the analysis is based on actual take-up. Doing so requires some assumptions, whereas the ITT estimates rely on only random assignment to ensure that the treatment and control groups are similar.[29] For example, Weathers and Stegman (2012) used two-stage least squares to analyze the impact of the AB demonstration on those who participated. Although the ITT approach requires fewer assumptions, sometimes it is important to learn about the impact on those who actually receive the intervention in addition to learning about impacts on those offered the intervention. SSA should consider computation of TOT estimates for future evaluations. They are relatively straightforward to do.

---

[27] An important technical topic that we do not address in detail here is correct estimation of standard errors in evaluations. Failure to take account of clustering, for example, can lead to underestimates of standard errors and incorrectly rejecting the null hypothesis of no impact. Although most of the reviewed SSA evaluations did not discuss the use of robust estimates of standard errors, the BOND evaluation is a notable exception (see Gubits et al. 2018).

[28] Some of the evaluations involved situations in which departures from the standard analytical techniques were warranted. BOND Stage 1 used a random effects estimator to generate externally valid hypothesis tests. In addition, as discussed further in the next section, the BOND evaluation adjusted the standard errors of the impact estimates to account for the design that was used. The MHTS evaluation also included major use of other statistical approaches to deal with specific issues, approaches that have rarely been used in evaluations of social programs. For example, the MHTS evaluation used negative binomials to estimate impacts when the outcome was a count variable that tended to mass at zero (e.g., number of months employed), ordered logit when the outcome was an ordered ordinal variable, and an analogue of the Wilcoxon test when the outcome variable was assumed to have a non-normal distribution.

[29] See, for example, Bloom (1984). The key assumption for Bloom's adjustment is that the treatment has no impact on those in the treatment group who do not receive the treatment. Also see Heckman, Smith, and Taber (1998).

Many of the evaluations made use of weighted regressions, rather than ordinary least squares, often to account for observations missed in surveys (discussed next). Although weighting is always required when making inferences about descriptive statistics, Solon, Haider, and Wooldridge (2015) suggest that there is considerable controversy about the use of weighting in estimating causal effects. This chapter is not the place to settle the disagreement, but we concur with them that "in situations in which you might be inclined to weight, it often is useful to report both weighted and unweighted estimates and to discuss what the contrast implies for the interpretation of the results" (314).

Treatment of Missing Data and Missing Observations

There are two types of missing data: unit and item nonresponse. Unit nonresponse occurs when an entire record is missing, such as when an individual does not respond to a survey. Item nonresponse occurs when only some of the variables for a given individual are unavailable. A common approach for unit missing data is weighting; a common approach for item missing data is to impute their values, often by using the means for those study participants for whom the data are available (see Puma et al. 2009).

The SSA evaluations often followed these missing data procedures, although some did not. For example, the evaluation of SPI simply excluded individuals from some analyses when there was missing data; in addition, it excluded about 2 percent of the sample because their earnings or hours appeared implausibly large.[30]

Unit and item missing data problems are usually less common in administrative data than in survey data. However, the non-experimental evaluation of HOPE relied on administrative data collected by programs serving persons with disabilities experiencing homelessness, and it suffered from both unit and item missing data: there were numerous missing forms, as well as missing items on the forms the evaluators did receive. Neither weighting nor imputation appears to have been implemented to treat these problems.

Another example of missing data is caused by withdrawals. For example, because the POD evaluation sample is restricted to volunteers, as in other demonstration

[30] In an unusual approach, the evaluation of YTD used an imputation procedure when the value of an outcome measure was missing and the measure was conditional on another outcome (e.g., earnings on employment status). Although this procedure introduces some uncertainty in interpreting the impact estimates, the evaluators state: "Impact estimates for outcomes with conditionally missing data would be biased if we did not adjust for missing information. However, when we calculated the biased impact estimates by dropping observations with missing outcome information, we found results very similar to those of the imputation procedure…. The similarity of the findings is not surprising, given the relatively small share of observations with missing outcome information" (Fraker et al. 2014, A.6).

Had missing outcome information been greater and the findings dissimilar, it is not apparent which set of results would be more acceptable.

programs involving volunteers, they are free to withdraw from the evaluation at any time. As explained earlier, members of the treatment group have an incentive to withdraw if they enter the TWP or the Grace Period, because entering causes them to be worse off than they would be under existing SSDI rules. Members of the control group did not have similar incentives to withdraw. Early in that demonstration, 4 percent of the treatment sample withdrew from POD, and virtually none of the control sample withdrew. The most common reason for withdrawing given by the treatment group was having earnings in the range in which their incomes would diminish (Hock, Wittenburg, and Levere 2020).[31]

Addressing the Multiple Hypothesis Testing Issue

Many of the SSA evaluations look at multiple outcomes; for example, employment, earnings, SSDI and SSI benefits, and physical and mental health. Moreover, they often use more than one measure of an outcome and more than one year of data. In addition, multiple treatment arms also result in multiple tests of hypotheses. For example, with two treatment arms, there are three comparisons: the two treatments with each other and each with the control group.

When multiple analyses are conducted, the probability of experiencing a "false positive"—meaning the null hypothesis of no impact is erroneously rejected—increases rapidly as the number of hypotheses tested increases. Schochet (2009) illustrates this problem by noting that if the Type I error rate is set at $\alpha = .05$, the probability of falsely rejecting the null hypothesis is 5 percent for each test, but "if all null hypotheses are true, the chance of finding at least one spurious impact is 23 percent if 5 independent tests are conducted, 64 percent for 20 tests, and 92 percent for 50 tests" (540).

Evaluators use several approaches to adjust calculations of statistical significance when multiple hypotheses are tested so that a statistically significant impact finding that could be due to chance does not get uncalled-for attention. Schochet (2009) reviews the procedures often used to deal with the multiple hypothesis problem, and he suggests identifying the most important hypotheses as "confirmatory" in advance of the empirical work and then considering all other hypotheses as "exploratory," where causal claims are not made.

Two of SSA's evaluations have considered the multiple hypothesis problem. The BOND evaluation identified earnings and SSDI benefit receipt as the two confirmatory

---

[31] In computing impacts, those who have withdrawn should probably be included in the sample used for estimation. This can be seen by considering POD's impact on earnings. Because earnings among the treatment group members who withdrew are likely greater than earnings among those who did not withdraw (Hock et al. 2020), dropping withdrawers from the sample would reduce the average earnings of the treatment group relative to the control group, causing the estimated impact on earnings to be biased downward. Note, however, that the unbiased impact estimate would pertain only to the intervention as it actually operated in the demonstration, not if POD is implemented nationally and withdrawals are not permitted.

outcomes, and the authors adjusted the statistical significance accordingly.[32] The YTD evaluation first defined five "research domains," each consisting of a different type of outcome (paid employment and earnings, total income from earnings and benefits, participation in productive activities such as employment and education/training, contact with the justice system, and self-determination as measured by an index). The evaluation then assigned one primary outcome to each of four domains and two outcomes to the fifth domain; it also examined secondary outcomes.

Evaluators disagree on when multiple hypothesis adjustments are required and on which adjustment should be used. Evaluators of SSA demonstrations and programs should be familiar with the issues, and they should consider the suggestion in Schochet (2009) to specify which hypotheses are considered confirmatory in advance of impact estimation.

Pooling across Sites

Most of the SSA evaluations we reviewed took place at multiple sites, and a decision had to be made on whether to analyze the sites separately or to pool data collected across sites in a single analysis. The exceptions were two evaluations that were conducted nationally—Ticket to Work and the Nudging Timely Wage Reporting experiment. In each of these two evaluations, an identical intervention was implemented across the country, and all data were pooled in each analysis.

There are rationales for both pooling across sites and not pooling. If the samples are large enough at each site, both strategies can be pursued. The primary rationale for pooling is that pooling increases the sample size, permitting estimates of the overall impact with greater precision and sometimes providing enough data to estimate subgroup impacts with sufficient precision. Pooling is the appropriate strategy if there is a uniform treatment (one intervention) and the target groups are the same across sites. Pooling is not appropriate if the treatments, the target groups, or both vary among sites and the intent of the evaluation is to determine the effectiveness of each intervention on each target group.

Most of the SSA demonstrations involved implementing an intervention (or similar interventions) for the same general population, and their evaluations pooled data across the demonstration's sites. We next describe the exceptions and variations.

PROMISE. The PROMISE set of six demonstrations included five state sites plus a consortium of six states. The population served was similar across the six sites, but the interventions varied somewhat. The impact evaluations were conducted separately for each site.

---

[32] Although many adjustment methods exist, this report used the Westfall-Young stepdown method, described by Westfall and Young (1993). A good explanation of the approach and how it was applied to an evaluation of a healthy marriage program is provided by Lowenstein et al. (2014).

YTD. The YTD included six sites, and its impact evaluations were conducted separately by site with no pooled impact analysis. Fraker et al. (2014) note that although the sites followed the same basic approach, there were meaningful differences among the sites: "All of these projects included the required components…but they took unique approaches to implementing them. The projects differed greatly in their organizational structures and the geographic and population sizes of their service delivery areas" (8). In particular, implementation at the second set of three sites differed in some ways from that at the initial three sites.

Project NetWork. Project NetWork tested four distinct models of delivering the intervention in two states each. Most of its impact evaluations were based on a pooled analysis, but Kornfeld et al. (1999) summarize the results by service model and provide details of the analysis by model in an appendix. Kornfeld and Rupp (2000) also summarize the findings by model.

DMIE. The evaluation that best exemplifies the case for separate site evaluations is DMIE. In each of four states, its evaluation selected a target group with specific disabilities, including mental health, selected mental and physical health disabilities, and diabetes. Whalen et al. (2012) conducted most of the impact analyses separately for each state, but they also pooled some analyses for two states because "the two states had similar participants with overlapping characteristics" (16).

In general, the SSA evaluations we examined appeared to weigh the pros and cons of pooling across sites. When the target groups and interventions were similar, the sites were pooled; when there were major differences, sites were analyzed separately; and when both approaches offered different benefits, both approaches were used.

Length of Follow-Up

The follow-up periods for the SSA evaluations vary. Some of the evaluations focused on short-term outcomes, such as the outcome of the application process or reporting earnings for a yearly period to SSA. These evaluations tended to have very short follow-up periods.

Nudging Timely Wage Reporting. This experiment tested four approaches for encouraging SSI recipients to report changes in their annual earnings. The intervention was very inexpensive and aimed to affect behavior for a maximum of only eight months, so a longer follow-up period was not needed. Also new notices are issued each year, so a longer follow-up would not be meaningful.

BEST. This was a proof-of-concept study, where the goal was to see whether applicants for SSI and SSDI experiencing homelessness could be processed more quickly when they received alternative services. Because there was no control or comparison group, the immediate outcomes were compared to outcomes for other applicants. Presumably, if SSA decides to conduct a rigorous evaluation of a program like BEST, follow-up periods similar to those used in other SSA evaluations would be used.

HSPD. Short-term follow-up was important in the HSPD evaluation, but longer-term follow-up could also be important. In HSPD, applicants experiencing homelessness who express symptoms of schizophrenia were provided with special services intended to speed up the SSI application process and improve the timeliness of benefit receipt; thus, the short-term outcomes were considered key in the evaluation, although longer-term outcomes were also of interest.

Demonstrations intended to have long-term impacts on employment, earnings, and receipt of SSI or SSDI generally had longer follow-up periods, and many of the evaluations included multiple follow-up periods. Several of the evaluations tracked outcomes at one year after random assignment or at completion of services (AB, DMIE, HOPE), but follow-up periods of two or three years were more common (BOPD, MHTS, DMIE for some participants, MHTS, POD, Project NetWork, PROMISE, TETD, YTD). The longest follow-up periods were four years for Ticket to Work and Phase 2 of BOND and five years for PROMISE and Phase 1 of BOND. Although the AB final evaluation report was based on only a one-year follow-up, the follow-up has already been extended for 3 years and may be further extended for 11 years. As previously mentioned, the follow-up for YTD has already been extended between five and seven years, with plans to extend it considerably further.

How long should follow-up periods be? There is no universal answer. The optimal period depends on how long the demonstration might anticipate benefits to last based on theory, prior experience, and evidence from earlier follow-ups. As discussed earlier, many of the outcomes associated with evaluations of SSA interventions can be captured by administrative data maintained by SSA—employment, earnings, SSI and SSDI benefit receipt, and death. If these are the primary outcomes of interest, long-term follow-ups can be conducted at a relatively low cost, at least as compared to evaluations that involve surveys.

Cost is not the only consideration in determining the follow-up period, however. If a program appears to have no initial impact, is it reasonable to assume there might be a "sleeper" impact where benefits occur a few years later? (See, e.g., Chetty et al. [2016]). More likely, if there are initial benefits in the form of increased earnings, how long should the follow-up be? In the employment and training field, evaluation of the Job Corps provides an important caution regarding extrapolating earnings gains. In a four-year follow-up, the program had strong earnings gains through the 48 months following random assignment (McConnell and Glazerman 2001). The evaluators projected that the earnings gains would be sustained. As a result, in their cost-benefit analysis, they estimated the present value of earnings gains after the observation period to be more than $27,000. In a later report, Schochet, Burghardt, and McConnell (2006) concluded that "according to the administrative records data, the estimated [earnings] impacts in years 5 to 10 for the full sample are all near zero and none are statistically significant" (3). Because earnings impacts were not sustained, Schochet et al. reversed the earlier conclusions: "Because overall earnings gains do not persist, the benefits to society of Job Corps are smaller than the substantial program costs" (3). The longer

time horizon revealed that a program that appeared to have social benefits that exceeded its costs in the short run did not in fact produce net social benefits because the benefits lasted only for five years.

The Job Corps results might not apply to the SSA demonstrations, but the point is that without a long enough follow-up period, policymakers must rely on extrapolating short-term findings. The implication is that for programs that appear to produce net social benefits and can use administrative data to track key outcomes, follow-ups should be conducted until projections are not needed to determine whether the present value of the program's benefits exceeds its costs.

A related issue is the amount of time over which a policy is tested. As discussed earlier, the evaluations of many demonstrations ideally should estimate the impacts of a permanent change in a policy, such as the reduction of the benefit reduction rate in BOND. If participants believe that a change in policies is permanent, how long the new rules are in effect is unimportant because participants will behave as if the new rules are permanent. If, however, participants are not sure a policy change is permanent—the reduction in the BRR in BOND was temporary, for example—they might not behave the way they would if it were permanent. The same general phenomenon arises in health insurance and income maintenance demonstrations. One way to determine whether the duration of a change affects impacts is to have treatment arms in which the change continues for different lengths of time. For example, in the Seattle and Denver Income Maintenance Experiment, one arm ran three years and the other arm ran five years. Comparing the two arms provided some indication of whether duration affected the response to the treatment (Burtless and Greenberg 1982).

Efficacy versus Efficiency

In discussing demonstration projects, the literature in public health distinguishes between efficacy trials and efficiency trials. *Efficacy trials* test the optimum implementation of an intervention, often at a small scale. Efficacy trials are conducted when, for example, programs are evaluated in the sites that are most likely to administer a treatment successfully, the individuals selected into treatment are those most likely to benefit from the treatment, the program was optimized for the conditions existing in the selected sites, or intensive technical assistance that would not exist in an ongoing program is provided to the sites (see Banerjee et al. [2017] for a discussion). Ideally, but not always in practice, *effectiveness trials* follow efficacy trials, when evaluations consider the program in a "real-world" setting, often increasing the scale of operations. This distinction is important because if an efficacy trail is conducted but an efficiency trial is not, the information available for launching the evaluated intervention as an ongoing program could be limited and possibly misleading.

The MHTS is an interesting example of an efficacy trial. The study sites were selected on the basis of their ability to deliver a complex of intervention services, which included supported employment, systematic medication management,

behavioral health and related services, prescription medicine, and comprehensive insurance. Fidelity to the intervention model was exceptionally rigorously tested and technical support was provided to sites that deviated from the model. Two of the original 23 sites ceased recruitment and enrollment activities in the first year of the evaluation because of internal operation issues (Frey et al. 2011).

The Role of Process Analysis

In evaluating an intervention, it is important to determine how it actually operates. For example, is it delivered in the manner intended by those who designed it? Do participants in the delivery program receive the intended services? Would they receive the same or similar services without the program? Are different subgroups of participants treated differently? Interpreting impact estimates requires answers to such questions. The purpose of process analysis is to provide the answers. In this subsection, we briefly discuss three overlapping types of process analysis: studies of how well the intervention is implemented and communicated to those receiving it; analyses of participation in the intervention program; and studies of fidelity to the intervention model.[33]

Implementing and Communicating the Intervention

One of the major roles of process analysis is to determine the ways the intervention—and components of it—are implemented, how quickly they are implemented, whether they are implemented as intended, and whether individuals eligible to receive the intervention understand it. For example, the process analysis conducted in evaluating the SPI demonstration included descriptions of the processes used in each of the four states to implement the waivers tested in the demonstration. It also included assessments from SSA field and regional office staff regarding waiver implementation and the ways in which the waiver processes affected other SSA operations, such as reducing overpayments. Implementation analysis, which is the most frequently conducted type of process analysis, commonly involves reviews of relevant available written materials and interviews; focus groups; or surveys of staff running the intervention program, individuals eligible for the intervention, or both groups.

The three SSA demonstrations that tested changes in the SSDI BRR (BOND, BOPD, and POD) illustrate the usefulness of process analyses in interpreting findings from impact analyses. For example, there was indication in all three studies that the treatment groups had difficulty understanding the changes to SSDI rules, which were complex, and especially complex in POD. This raises the question of whether the

[33] Details on findings from process analyses of the SSA demonstrations, with particular emphasis on recruitment and enrollment into the demonstrations and program delivery of services, can be found in Chapter 9 in this volume.

behavior responses to the intervention were suppressed by this lack of understanding, thereby muting the impact estimates, and whether similar muting would occur with a permanent policy change that allowed time for a greater understanding. In addition, the BOND final report concluded that there was less outreach in Stage 1 to inform beneficiaries about the offset than there likely would be if the tested rules were implemented permanently (Gubits et al. 2018a/b).

Participation in the Intervention

Participation analysis, a subcategory of process analysis, involves determining the percentage of the treatment group, and sometimes the percentage of the control or comparison group, that actually participates in the intervention being tested (e.g., that receives services). In addition, the characteristics of those who participate might be compared to those who do not. In the case of financial incentives, such as those provided by BOND, the process analysis also might include determining the percentage of the treatment group whose SSDI or SSI benefits are affected.

Participation analysis is usually performed with data collected from surveys, available from management information systems, or sometimes from SSA administrative records. For example, using administrative data, the SPI evaluation determined what percentage of SSI recipients who were offered each of the four tested waivers actually used them. Similarly, the evaluation of BOND used SSA administrative records to examine the fraction of Stage 1 and Stage 2 treatment group members who used the financial incentive (i.e., the offset). When programs and policies involve multiple components (e.g., training and job placement), it is important to estimate participation in each program component. As previously mentioned, for instance, AB Plus provided health insurance and, in addition, treatment group members qualified for three different services that were accessed over the telephone. Using management information records, the evaluators computed distinct participation rates for the use participants made of the provided insurance plan and each of the telephone services (Michalopoulos et al. 2011). Finally, if some members of the control or comparison group receive services similar to the intervention's from non-program sources, then their participation rates in those services should also be determined.

Based on survey data, the evaluation of Project NetWork estimated participation rates for both treatment and control groups for 10 separate services, finding that participation rates were fairly small for most services and that rates were not much higher for treatment group members than for control group members (Kornfeld et al. 1999). Obviously, if there is little participation by treatment group members or little difference between treatment and control group participation rates, then impacts of the intervention on other outcomes are also likely to be small.

Fidelity to the Intervention

Unless there is reasonable fidelity to the program model of the intervention being evaluated, it is not possible to interpret impact estimates, regardless of whether they are favorable or unfavorable, because what generated them is unknown. Moreover, once a lack of fidelity is uncovered, technical assistance can be provided to correct the problem.

To the extent process studies determine whether an intervention was implemented as intended, they provide considerable information about fidelity. Sometimes, however, a further useful step is to develop an index to measure fidelity to the program model. One of the SSA demonstrations, the MHTS, did so. For this purpose, the evaluators used a 15-item measure, the IPS Fidelity Scale, where IPS (Individual Placement and Support) refers to the program model. The scale for each item ranged from a low of 1 (poor adherence to the model) to a high of 5 (close adherence to the model). The scale was administered annually by a designated team at all 23 of the study sites. Based on the results, the sites were provided feedback and, when needed, technical support (Frey et al. 2011). One potential use of a formal fidelity measure such as the IPS Fidelity Scale is that it can be incorporated into a multiple-site evaluation to see whether program impacts vary with fidelity score (see Greenberg, Meyer, and Wiseman 1994).

It is evident that developing and implementing a formal fidelity measure requires considerable resources, suggesting that doing so should be limited to complex interventions such as the MHTS intervention, which included clinical services. At a minimum, studies of program implementation and participation should almost always be part of an evaluation.

Role of Cost-Benefit Analysis[34]

Cost-benefit analysis (CBA) assesses the net present value of economic gains or losses from an intervention by comparing its benefits with its costs. It usually does this from the perspective of society as a whole and also often from the perspective of the groups that compose society. The cost-benefit analysis of BOND, for example, examines benefits and costs from the perspectives of four groups: SSDI beneficiaries, the Disability Trust Fund, the rest of government, and society as a whole (Gubits et al. 2018a/b). "Society as a whole" is simply the sum of the benefits received and the costs incurred by the first three groups and by non-beneficiaries. The benefits and costs

[34] In addition to conducting cost-benefit analyses as part of program evaluations, as Jesse Rothstein comments on this chapter, prospective CBAs can be useful in determining whether a proposed intervention is worth testing. By conducting a CBA before the demonstration, one can assess whether the impact required to achieve a positive net present value is feasible. Anticipated program impacts can sometimes be gauged by a literature review, meta-analysis, or microsimulation.

included in a CBA must be estimated in monetary terms such as dollars in order for them to be summed.

Six of the SSA evaluations we reviewed included CBAs as part of their evaluation plan. Some of these CBAs have been completed, and others are planned. In addition, a cost-effectiveness analysis, in which costs were monetized but benefits were not, was conducted in one evaluation (Nudging Timely Wage Reporting); and program operating costs were estimated in two evaluations (MHTS and AB).

Estimates of program operating costs are needed for budgetary purposes by agencies running a demonstration. However, if services offered by a program substitute for similar services available elsewhere, such an estimate might not be sufficient for CBA purposes. Stated a bit differently, estimates of operating costs are measures of gross costs, not net or incremental costs. It is, however, estimates of the net or incremental costs (which are usually obtained by comparing the costs of services received by a treatment group with the costs of similar services received by a control group) that are essential for cost-benefit analysis.

CBAs usually examine a much larger range of benefits and costs than impact analyses do. For example, in addition to increases in earnings and SSDI benefits, the CBA of BOND included estimates of the impacts of the policy change on fringe benefits; SSI payments; income, sales, and payroll taxes; work-related expenditures (e.g., child care and transportation); the costs of the Ticket to Work program and state Vocational Rehabilitation programs; economic distortions related to changes in the government's fiscal position; and time available outside of work (Gubits et al. 2018a/b).

Many of the key benefits used in CBAs, such as program or policy impacts on earnings and transfer payment receipts, are obtained directly from impact analyses. Other benefits, such as fringe benefits and tax payments, are derived indirectly from the impact estimates. For example, an estimate of BOND's impact on fringe benefits was computed as a multiple of the estimate of BOND's impact on earnings. Thus, CBAs are highly dependent on impact analyses. The other major input into CBAs, net program operating costs, is typically obtained from a separate cost study.

As is evident, if a CBA is to be conducted, evaluation designs must include plans for collecting data on both the key outcome measures and the necessary cost information. Because cost-benefit analysis incorporates multiple impacts that could work in opposite directions, the net benefits of an intervention can demonstrate that an intervention is worthwhile even if its impacts on earnings and transfer benefits are negligible.

In principle, the impacts of interventions can persist for many years. For example, impacts on earnings could potentially continue until the members of a treatment group retire. Benefits and costs would ideally be included in a CBA for every year for which they continue to exist. Because SSA administrative data follow individuals over time, they are ideal data for this purpose. However, policymakers usually want evaluation findings as soon as possible, rather than waiting until the members of a research sample

retire. As a result, a compromise involving projecting effects is often made. For example, the CBA of Project NetWork is based on observing impacts for two years for part of the sample and three years for the remaining sample and then projecting impacts for an additional two or three years (Kornfeld et al. 1999). The evaluation of YTD has conducted a "preliminary" CBA based on only 3 years of data (Honeycutt, Morris, and Fraker 2014), but SSA plans to internally conduct a future CBA based on a much longer observation period, possibly up to 25 years.

AREAS FOR FURTHER EXPLORATION

This section discusses topics that received little or no mention in the final reports of the SSA evaluations we reviewed. These include design innovations that SSA might consider in future evaluations. The section also considers potentially important program impacts that have seldom been estimated in SSA evaluations because doing so is difficult.

Alternative Experimental Designs

The essence of experimental evaluations is the use of random assignment as the method of allocating individuals to treatment and control groups. In this subsection, we introduce variations in the way that random assignment can be carried out. The simplest approach is for each individual to have the same probability of being assigned to either treatment status; if there is a single treatment and a control group, for instance, then each study enrollee would have a 50 percent chance of being assigned to either status.

There are several reasons why the probability of assignment might not be uniform.[35] First, if the budget for the evaluation includes the cost of the intervention being evaluated, then treatment cases require much more expense than control cases do. If the control group is larger than the treatment group, more individuals can be included in the evaluation, and treatment impacts can be estimated more precisely. Second, if individual sites must volunteer to participate in the evaluation, then they could be more agreeable if only a small portion of the research sample will be assigned to the control group and denied services (assuming the treatment adds desirable services).

If there are two treatment groups and a control group, then the issue of what assignment ratio to use becomes more complex, depending in large part on how the data will be analyzed. If the most important hypotheses involve combining the treatment groups, as is sometimes the case (e.g., when the hypothesis of most interest is whether receiving any of the treatment services has an impact, rather than assessing the impacts of alternative treatments), then the optimum design will assign fewer cases

[35] BOND stage 1 and YTD are examples of demonstrations where probabilities of assignment to treatment and control status were not equal.

in each treatment group, than if the most important hypotheses concern the relative impacts of the alternative treatments.

Clustered Designs

Hussey and Hughes (2007) note that "cluster (or community, or group) randomized trials (CRT) are distinguished by the fact that individuals are randomized in groups rather than individually" (182). They observe that "cluster designs may be chosen because the intervention can only be administered on a community-wide scale, or to minimize contamination, or for other logistic, financial, or ethical reasons" (182). The major drawback of cluster designs is that they usually lack sufficient statistical power because they generally have too few sites.

Stepped-Wedge Designs with a Staggered Rollout

Stepped-wedge designs are a type of "staggered introduction design," where initially none of the clusters has the intervention, then over time, the intervention is gradually introduced. In this way, late implementing clusters serve as comparison groups for early implementers (Peck 2020, 40). Hussey and Hughes (2007) define the stepped-wedge design as follows:

> A stepped-wedge design is a type of crossover design in which different clusters cross over (switch treatments) at different time points. In addition, the clusters cross over in one direction only—typically, from control to intervention. The first time point usually corresponds to a baseline measurement where none of the clusters receive the intervention of interest. At subsequent time points, clusters initiate the intervention of interest and the response to the intervention is measured. More than one cluster may start the intervention at a time point, but the time at which a cluster begins the intervention is randomized. (183)

The stepped-wedge design can be a useful way to evaluate an intervention that eventually will be provided to the entire population, particularly when it could be considered unethical to withhold the intervention for an extended period.

There are, however, some aspects of this design that limit its utility. First, the clusters in treatment and control status might not be similar in characteristics that affect the outcomes of interest. If so, differences in the outcomes of treated and untreated clusters can result from baseline differences, rather than from presence of the treatment. Depending on the number of clusters, this problem can be mitigated to some extent by randomizing when the treatment is implemented in each cluster. The Ticket to Work evaluation was based on a variant of a randomized stepped-wedge design. All SSI and SSDI participants were entitled to receive a ticket, but the tickets

were allocated monthly based on the last digit of the Social Security number, which is equivalent to random allocation (see Livermore et al. 2013).

Second, the stepped-wedge design is more valuable for measuring short-term impacts than long-term impacts. Suppose, for example, the comparison clusters transition to treatment status on a monthly basis, and the evaluators are interested in the impact of the intervention on earnings, say, 10 years later. At the end of 10 years, the evaluation would not be able to observe groups with and without the treatment; it could only observe groups that had the treatment (say) 10 years ago and compare them to individuals from groups that had the treatment 9 years ago. For an intervention such as job search assistance, where the impact is likely to take place immediately after the intervention and then decay to zero fairly rapidly, a stepped-wedge design can be adequate. For a potentially long-lasting intervention, such as occupational training, the design is less useful. The timing of the rollout should be chosen to align with information needs.

Adaptive Designs

By adaptive designs, we mean modifications in the evaluation design as a result of preliminary evaluation findings. One example is early-stopping designs in which minimum target impact values are set prior to beginning a demonstration. If the estimated impacts fail to meet these targets, the demonstration and its evaluation could then cease.[36] Other adaptive designs involve modifying a treatment to make it more attractive to the target population, improving communication about the treatment, and augmenting the size of the sample or modifying the randomization procedure to increase the chances of obtaining a statistically significant finding. Chow and Chang (2012) provide a comprehensive summary of adaptive designs in clinical health trials.

[36] Although early stopping can result in considerable resource savings, it should be used with caution because early findings from an experimental evaluation in the social policy area can be highly misleading. For example, the United Kingdom's Employment Retention and Advancement demonstration's early impacts on earnings appeared very promising for unemployed single mothers receiving welfare and more modest for long-term unemployed men. However, these impacts faded for the former group but were sustained for the latter group. As a result, a cost-benefit analysis found positive net present values for the unemployed men, but not for the single mothers (Hendra et al. 2011). This finding was unanticipated by those involved in the evaluation. Important ethical issues can also be raised by early stopping. During the 33 months the Employment Retention and Advancement demonstration was scheduled to continue, participants were promised a substantial cash incentive three times a year if they worked at least 30 hours a week for 13 out of every 17 weeks. If the demonstration had been prematurely terminated for men in the treatment group based on those early findings, then bonuses would have been lost to the men expecting them.

Factorial Designs

Factorial designs are the natural next step beyond a multi-armed experimental evaluation. Peck (2020) defines a factorial design as one that "varies two (or more) treatment dimensions or factors, randomizing to each individually and to both together. If the levels of each factor include 'absence' or 'presence,' then the absence of both factors represents a status quo control group" (78). Factors can either vary in dosage or simply be present or absent.

As an example, consider a modification to the SSDI program where SSA wants to test two variations of the reduction for earned income (the current SSDI cash cliff versus a 50 percent BRR) and two variations in the threshold at which benefits currently cease (the current threshold versus a higher threshold that is twice as large as the current threshold). In a factorial design, participants are assigned to one of the possible combinations of the factors. In the situation described above, these would be (1) the cash cliff and the current threshold; (2) the cash cliff and the higher threshold; (3) 50 percent reduction of the benefit for each dollar of earned income and the current threshold; and (4) 50 percent reduction of the benefit for each dollar of earned income and the higher threshold. If the factorial design is applied to an ongoing program, one of the factor combinations is the current design (the first design in the example), which is a type of control group. In a training program demonstration, a control condition can be included where all factors are set to "no services."

Factorial designs have been used in random assignment evaluations to evaluate health insurance programs and welfare policies. In the example above, the two factors are the BRR and the threshold at which the reduction in benefits is applied. The primary advantage of factorial designs is that they can be used to estimate the impacts of each factor separately and every combination of the factors.[37] The primary disadvantage of factorial designs is that to estimate all treatment combinations, the required sample size increases, as does the cost of the demonstration.

Other Experimental Designs

There are many variations on how random assignment can be implemented in an evaluation. Examples are provided in Peck (2020) and Orr (1999), but these sources are not exhaustive. The best design for a specific evaluation will depend on which hypotheses are most important to test, cost limitations, ethical considerations, and practicalities.

Seldom-Estimated Impacts

Interventions that include policy and program changes can affect outcomes in numerous ways. This subsection discusses some impacts that are potentially important

[37] Peck (2020) notes that a 2 × 2 factorial design can be used to test eight hypotheses.

under some circumstances but difficult to estimate. As a result, they were seldom addressed in the final reports of the SSA evaluations we reviewed.

General Equilibrium Effects

SSA policies and programs that affect labor market behavior such as employment placement and training programs (e.g., MHTS, TETD, Ticket to Work, YTD) and policies that change financial work incentives (e.g., BOND, BOPD, POD) can have effects on the well-being of individuals who themselves do not receive SSDI or SSI, and because of this, on the general economy. We consider three types of these effects next (see Greenberg et al. [2011] for a fuller treatment of the issues).[38]

Displacement Effects

Job training programs or financial work incentives policies, if successful, can increase competition for available jobs. As a result, individuals who are directly affected can end up in jobs that would otherwise have been held by those not directly affected by the programs or policies (Johnson 1979; Schiller 1973). If so, the earnings of the latter are less than they otherwise would be, and consequently the net benefits of the programs or policies are less than otherwise would be the case. For example, as shown in Exhibits 1.6 and 1.7 in Chapter 1 in this volume, the TETD program had modest but positive impacts on the employment and earnings of the SSI recipients with intellectual disability who received the services offered by the program. It is possible that in the absence of TETD, persons who were not receiving SSI would have occupied these positions.

The importance of displacement effects partially depends on the number of existing job vacancies. The fewer the number of job vacancies, the more difficult it is for unemployed individuals who are *indirectly* affected by the programs or policies to find jobs that are alternatives to the jobs taken by the unemployed individuals who are *directly* affected. As a result, the latter have "displaced" the former in the job market. This suggests that the size of the displacement effect is likely to reflect the state of the relevant local labor markets. However, even if there is high unemployment and substantial displacement, it is unlikely to be permanent. If the economy is expanding, the displacement effects should diminish over time, as job opportunities open and absorb those who were displaced.

As a result, the displacement effect is likely to be more important in the short run than in the long run. Moreover, as emphasized by Johnson (1979) and Katz (1994), if

[38] A fourth type is "multiplier effects," which refer to the possibility that SSA interventions might stimulate the economy through employment, subsequent consumption, and so on. Multiplier effects are germane only when unemployment is substantial. In general, multiplier effects are probably best ignored in evaluations of training programs. This is because any multiplier effect that results from training program expenditures is likely to offset multiplier effects that would have occurred had the same funds been used for an alternative purpose.

training programs can impart skills that allow trainees to leave slack occupational labor markets for tight ones, then programs decrease the competition for job vacancies in the slack markets, thereby making it easier for those in the slack labor markets who are ineligible for the program to find jobs. Such a possibility could produce a result that is the exact opposite of a displacement effect—total employment could increase by more than the number of persons who are trained.

It is rarely possible to estimate the size of displacement effects as part of an evaluation of a specific program or policy (an exception is Crepon et al. [2013]). That being the case, whenever favorable impacts on employment are found in an evaluation, we suggest that displacement should be mentioned in the evaluation report as a potential unmeasured effect of uncertain size. This is especially relevant in the context of cost-benefit studies, such as the one conducted as part of the evaluation of TETD, where displacement should be appropriately viewed as a negative benefit from the perspective of society as a whole. The state of the labor market in the evaluation sites should be considered in this discussion because displacement effects will likely be larger where unemployment is higher, and they will diminish over time if the economy is expanding. For example, the unemployment rate was relatively low at the time the TETD demonstration was run, suggesting that displacement may have been modest.

Fiscal Substitution Effects

Akin to displacement effects, a "fiscal substitution" effect (Johnson and Tomola 1977) can occur when the government provides employment subsidies or directly places targeted disadvantaged individuals into jobs at government agencies or non-profit institutions. For example, some YTD sites paid subsidies to private sector employers to hire members of specific disadvantaged target groups. Under such programs, the targeted group members might be hired instead of, or even replace, group members who are not targeted (subsidized) and so are more expensive for employers to hire. An example is when a local government uses individuals paid for by the federal government under a jobs program rather than hiring employees that the locality must pay for (Johnson and Tomola 1977). This is a concern because although employment among the target group could increase, to the extent fiscal substitution occurs, this favorable effect is offset by decreases in employment, among others.

Research on fiscal substitution effects suggests that they are often large, sometimes finding that half or more of any gain in earnings by program participants is offset through loss of earning by those substituted for (see the review of the empirical literature by Greenberg et al. [2011]). As with displacement effects, the implications for interpreting evaluation results of fiscal substitution effects should be mentioned in evaluation reports on programs that can potentially cause them—for example, the YTD sites that paid subsidies to private sector employers.

Equilibrium Wage Effects

If those affected by training programs or financial work incentives search harder for jobs or if their job skills increase—and, as a result the amount they work is greater than it otherwise would have been—then the resulting increase in labor supplied will tend to put downward pressure on equilibrium wages within the labor markets in which they work. As a result, workers who are employed in those same labor markets might receive lower wages than they otherwise would, a consequence that program evaluations are unlikely to capture. Most, but not all, of the empirical literature concludes that such effects are typically fairly modest (see Greenberg et al. 2011). Although most SSA programs or policies seem unlikely to bring about large equilibrium wage effects, we believe that future evaluations would do well to consider whether these effects are likely to have occurred. For example, one can consider whether the program accounted for a relatively large proportion of the supply population in specific labor markets.

Entry Effects

If a job placement or training program or a financial work incentives policy for SSDI beneficiaries or SSI recipients is perceived as attractive, but is available only to those on SSDI or SSI, some individuals might apply for SSDI or SSI benefits in order to access the program or policy (an "entry effect"). In contrast, if a program or policy is viewed as unattractive (e.g., a mandatory training program), some individuals who might otherwise have taken up SSDI or SSI could decide not to do so. The latter effect on entry is sometimes known as a "deterrent effect." Deterrent effects seem likely to be more important than entry effects for SSDI and SSI programs, because entry into these programs is difficult. For example, qualifying for benefits is contingent on a medical examination and on not having earnings for at least five months and often longer.

Moffitt (1992a, 1996), who first introduced the topic to the evaluation literature, argues that both entry effects and deterrent effects could be substantial. Entry effects will continue to occur over the long run and are unlikely to be fully observed in evaluations of programs and policies being tested as a demonstration. By definition, deterrent effects keep individuals from volunteering for a program or cause them to withdraw if they have already volunteered. In the case of mandatory job training in exchange for transfer benefits, for example, some individuals might withdraw from the benefits program or not enroll in the first place. Though evaluators would be able to observe withdrawals, they cannot observe individuals who do not enroll in a program such as SSDI or SSI.

Not surprisingly, empirical evidence about the magnitude of entry effects is quite limited. Most of what does exist pertains to welfare-to-work programs in the United States and Canada (Greenberg et al. 2011). Research on program entry effects is usually conducted separately from the evaluations of these programs and based on

aggregated data. Most of the findings are consistent with what might be anticipated: mandatory welfare-to-work programs consistently seemed to modestly discourage entry by making it more burdensome to receive welfare, whereas there is some evidence (although not as consistent) that voluntary programs tended to encourage modest entry onto the welfare rolls by providing services that might otherwise be difficult to obtain. The modesty of these estimates possibly suggests that entry and deterrent effects need not be considered a major issue in SSA evaluations.

Nonetheless, as discussed in Chapter 3, there was some concern prior to the BOND evaluation that the intervention might have an entry effect into SSDI as a result of the attractiveness of the benefit offset and that such an effect could not be measured by BOND's experimental design. As a result, although not ultimately taken up by SSA, several alternative designs for estimating BOND's effect on entry were proposed (Tuma 2001; Maestas, Mullen, and Zamarro 2010).

Program Component Effects

Most training programs consist of multiple components. A training program could offer help with searching for a job, counseling, basic education, more advanced education, Vocational Rehabilitation, on-the-job-training, classroom training, supportive services, and financial help in the event of emergencies. The Ticket to Work program, for example, allows SSDI beneficiaries and SSI recipients to use training and a variety of other services to assist themselves in obtaining employment. Even though few trainees will participate in all the components of a training program, many are likely to participate in more than one. Policymakers would, of course, like to know which components or sets of components are effective and which are not and the characteristics of the trainees for whom each component or component combination works best.

Learning about the relative effectiveness of various services is difficult. An obvious approach is to compare individuals who receive different combinations of services within a program. However, regardless of whether the services are selected by those running the program or by the program participants themselves, as in Ticket to Work, those receiving various services are likely to vary from one another in their labor market potential. For example, those receiving only help in job placement are likely much more job ready than those receiving basic education and Vocational Rehabilitation. This suggests that comparing labor market outcomes such as earnings to measure effectiveness is highly problematic. Another approach is to compare outcomes at program sites that emphasize different combinations of services. However, again, the client populations and local economic conditions could differ across sites, making it difficult to isolate the effects of the program design (see Barnow and Greenberg 2020).

Multi-armed experimental evaluations are probably the best way to learn about the relative effectiveness of alternative services or to isolate the relative impacts of components of a set of services that make up a multifaceted program. Factorial designs

offer the opportunity, as well. SSA has used multi-armed designs, but not factorial designs. Bell and Peck (2016b) describe a number of ways multiple arms, multistage randomization, and factorial designs can be used "to measure the contribution of specific features of interventions to overall impacts" (106). They also provide useful examples of when these designs have been used in practice. When they are not used, it could be necessary to use non-experimental methods to attempt to estimate the impacts of alternative components.

Site Representativeness

In the section "Major Evaluation Design Lessons," we discussed population-representativeness; the idea that the sample used in an evaluation of a demonstration project should ideally be representative of the individuals who would be eligible for the intervention being evaluated were it be rolled out nationally. The "Population-Representativeness" subsection above discussed two reasons why population-representativeness might not occur: the demonstration sites might not be representative of the target population; and, even within each demonstration site, the individuals affected by the intervention might differ from those affected were the program rolled out nationally. This subsection discusses how the first issue might be addressed.

Olsen et al. (2013) argue that most evaluations use purposive (i.e., convenience) samples of sites that are readily available, and that unless site impacts are identical across sites, impact estimates from such samples of sites are likely to be biased estimates of the impacts for the full population of interest. They offer several suggestions for coming closer to site representativeness than is often the case.

Site representativeness would be best accomplished by randomly selecting the sites from the full population of potential sites. The BOND evaluation is one example of when this was done. Olsen et al. (2013) make several suggestions to help approximate the random selection of sites when doing so is infeasible. One is to explore what characteristics make sites more likely to participate in a purposive study, and to compare impacts from these types of sites versus what would be obtained in a study in which sites were randomly selected. In addition, they suggest strategies that can be pursued to minimize the likelihood of refusal to participate in the study, such as providing incentives and passing laws requiring participation. Their third suggestion is to offer inducements to sites that initially refuse to participate and then compare the impacts of the original sample with the impacts of the sites that participate after additional recruitment efforts.

The final suggestion offered by Olsen et al. (2013) is to gather additional site characteristics and estimate the probability that various sites would participate and then use this information to develop weights for the analysis based on participation probabilities. They note that work on increasing external validity is at a formative

stage, but they believe evaluations will be more useful if external validity shortcomings are recognized and efforts are made to correct for the bias.[39]

CONCLUSIONS ABOUT EVALUATION DESIGN LESSONS

This chapter has examined 16 SSA evaluations that served the target populations of the SSDI and SSI programs. We focused on the design of the evaluations in order to provide strategies and lessons for future SSA evaluations. The evaluation designs are quite diverse. Most of the studies were experimental, but four were non-experimental and two of them were proof-of-concept studies that were not intended to provide impact estimates.

The evaluated interventions varied enormously. Three emphasized removal of the SSDI cash cliff threshold, one provided financial work incentives through waivers, three helped individuals apply for SSDI and SSI, one provided health insurance, one improved access to medical care and support services for individuals with disabilities not on SSDI or SSI, one sent letters to SSDI beneficiaries to nudge them to self-report their earnings, and six provided services intended to facilitate employment. The types of interventions that were evaluated strongly influenced the outcome measures that the evaluations emphasized, with earnings, employment, SSI and SSDI payments, health, and application speed and success playing important roles in different evaluations. Most of the evaluated interventions could involve only individuals who first volunteered, but three covered all SSDI beneficiaries who met certain criteria. In some of the SSA evaluations, but far from all, there were reasons to be concerned that they were not sufficiently population-representative.

Most of the evaluations assessed only a single treatment arm, but three examined two treatment arms, and one assessed four. Most of the SSA evaluations took place at multiple sites, and most of these pooled the findings across their sites, but a few did not. Most used SSA administrative data, and some also collected survey data. Almost all conducted a process analysis, although the methods used varied considerably; and about half also conducted a cost-benefit analysis or cost analysis.

Similar variation can be found in evaluations of programs and policies targeting other disadvantaged groups such as the unemployed and those participating in Temporary Assistance for Needy Families (TANF) and Supplemental Nutrition Assistance Program (SNAP) programs. What makes the SSA evaluations unique is that they target individuals with disabilities who either receive SSDI or SSI or are candidates to receive these benefits. As a result, most of the evaluations could use SSA administrative data. The SSA administrative data are arguably superior to

[39] There is some literature on manipulating results from an evaluation's sample to reflect the broader population of interest; this literature often makes use of post hoc propensity score methods (e.g., Stuart et al. 2011). Tipton (2013, 2014) and Tipton and Peck (2017) suggest a design approach for ensuring the generalizability from an evaluation's sample to a larger population.

administrative data from state UI programs, the data on which most evaluations involving other disadvantaged target groups have relied. Because the SSDI and SSI programs are difficult to enter, the SSA evaluations were probably also less subject to entry effects. Evaluations of interventions targeting the recipients of UI, TANF, and SNAP have typically been mandatory, whereas those focused on individuals with disabilities typically are not. Because the latter are voluntary, they are probably less subject to deterrent effects.

SSA has done an admirable job over the past nearly four decades in using demonstrations as a means to uncover the impacts of its potential policy changes. Indeed, the large majority of its demonstrations have involved experimental evaluations. The result is that a strong evidence base exists to inform decisions in this policy arena.

Our recommendation is that SSA continue to prioritize use of experimental evaluation designs. In this chapter's "Areas for Further Exploration" section, we suggested how the agency might push the envelope further.

Chapter 2

Comment

Jesse Rothstein
University of California, Berkeley

Burt Barnow and David Greenberg (in "Design of Social Security Administration Demonstration Evaluations") have done an excellent job summarizing the design of 16 evaluations conducted by the Social Security Administration (SSA) of demonstration programs involving Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). They methodically and thoroughly review how the different evaluations made choices around research design, statistical power, population-representativeness, data sources, missing data, and so on.

My comments here will focus on the interplay between the design of evaluations and the intended or expected use of the evaluation results in support of policy decisions. I focus on impact evaluations, typically randomized experiments, that infer the effect of a program on participants by comparing their outcomes to those of others exposed to a control condition.

I emphasize that my comments are not intended as criticism of SSA's past or current practice—overall, I am impressed at the care taken in the design and implementation of SSA's demonstration studies, many of which operated under externally imposed legal, logistical, or budgetary constraints. My comments are aimed primarily at policymakers interpreting the results of such constrained evaluations, and secondarily at evaluators, at SSA and elsewhere, who may in the future face design choices that could be informed by these considerations to better support the decisions that ultimately will depend on them.

WHAT TO EVALUATE?

A major question is what types of demonstrations to evaluate, and when in the policy development process it is appropriate to conduct a formal impact evaluation. Barnow and Greenberg distinguish *efficacy trials* from *effectiveness trials*, terms that I believe are borrowed from medical research. In Barnow and Greenberg's descriptions, efficacy trials "test the optimum implementation of an intervention, often at a small scale," whereas effectiveness trials "consider the program in a 'real-world' setting, often increasing the scale of operation." This is a useful distinction, and both types of trials are important. But they are not sufficient. These types of trials are appropriate primarily when we begin with a well-developed, carefully specified "intervention" that we want to study, for the purpose of deciding whether to implement it at a large scale, or perhaps to abandon it.

This is not the only value of policy demonstration and evaluation research. Another situation, arguably more common, is where policymakers have a theory about a potentially desirable change but are not sure whether the theory is correct or, if it is,

how to best use that theory to achieve desired outcomes. For example, policymakers might have a theory that some SSDI recipients are physically and mentally able to return to work but are prevented from doing so by the financial incentives built into the benefit structure. This theory, if correct, might support programmatic changes that reduce the rate at which benefits are reduced when earnings increase (as in the State Partnership Initiative demonstration [Kregel 2006b] or in BOND) or that allow participants to remain in the program even when earnings exceed the usual threshold (a variant of which is included in POD). But there are many potential programmatic changes that would accomplish this.

An efficacy trial would be appropriate if we had a single proposed change to consider—if the only decision to be made is whether to expand that exact change to the broader population or to abandon it, and there was no question about whether other potential changes might be better.[40] But often there are other decisions that we would like a demonstration to support—for example, whether we should further explore other similar changes, or look elsewhere for solutions to perceived problems. An efficacy trial is not designed for this.

This suggests that there is value in considering a third type of trial. Ludwig, Kling, and Mullainathan (2011) propose "mechanism experiments," where the goal is not to test a specific intervention as a program but to assess whether a hypothesized mechanism or theoretical channel is operative. One might use a mechanism experiment to test an intervention that would never be rolled out at a very large scale but that is well suited to assess the validity of a behavioral theory, with an idea that if the trial is successful then it could be used to support the design of a new intervention that exploits the same theory in a different way and that would be more realistic for large-scale implementation.

In the example of work incentives for SSDI recipients, a mechanism experiment might explore a very high powered incentive, such as a dramatically increased earnings disregard or a large wage subsidy, that would be too expensive to plausibly implement on a large scale but that would permit a clear test of the underlying theory. A version of this has been talked about as the "Ultimate Demonstration," which would allow SSDI beneficiaries to earn any amount without facing benefit reductions (see, e.g., Gubits et al. 2019). If the work incentives theory is correct, this high-powered treatment would surely yield sizeable impacts on beneficiary work. It could then be followed up with efficacy studies of lower-powered interventions, and then by efficiency studies. On the other hand, if the Ultimate Demonstration did not yield labor supply effects, we would have clear evidence that no incentive-based strategy is likely to work.

---

[40] In some cases, legislation may specify a particular policy change to be implemented and evaluated. Even here, this change can be thought of as an example of a family of potential changes to be assessed, rather than as the only change of interest; often, though not always, legislators may be interested in considering future implementation of another policy from the broad family, rather than just the specific policy specified for evaluation.

An advantage of adding a category of mechanism evaluations to the toolkit is that it might help to avoid category errors that are common in the policy use of program evaluation evidence. It is common to interpret a failed efficacy study as an indictment of the entire underlying theory rather than just of the specific program that was evaluated—in effect, treating it as a mechanism study though it was not designed as one.[41] But when the study considered only a single example, one not necessarily well crafted to test the mechanism, this conclusion may not be supported.

Indeed, some studies that are conceptualized as efficacy studies are really intended as mechanism studies, as the implicit intent is to assess not a specific intervention but a category of intervention. For example, Congress may specify a particular demonstration, but in fact be interested in exploring a possible direction for policy change rather than the specifics of the intervention to be evaluated. It is much better to recognize this explicitly. In some cases, this can support better study designs—for example, as in the Ultimate Demonstration, amplifying the "dosage" of the treatment to ensure that if the mechanism is operative, it will be found, even though such a high dosage would not be realistic in a larger-scale program. In other cases, legislation may not give SSA that flexibility, but policymakers may be able to more intelligently consider the generalizability of the results if they recognize that the study was a partial test of a mechanism rather than just a test of the efficacy of the particular intervention studied.

STUDY IMPLEMENTATION AND POLICY

Once a decision is made about exactly what intervention will be studied, there are several additional ways that demonstration practice can better reflect the potential policy uses of the study. I briefly review two here.

First, Barnow and Greenberg discuss the importance of including prospective power calculations in the design of evaluations. These are statistical calculations made at the outset of a study of the "minimum detectable effect" (MDE), the smallest true effect of the intervention that the evaluation would have a reasonable chance of being able to distinguish from zero. The goal is to avoid underpowered studies that do not generate precise enough effect estimates to support decisions.

I would argue that evaluators should—and indeed often do—go further, and include not just MDE estimates but prospective cost-benefit analyses or threshold analyses that identify how large the effect of the intervention would need to be for the program to be considered successful. Design studies should make clear how the MDE relates to the threshold analysis, ideally justifying the chosen MDE as a policy-relevant impact. This would help guard against a frequent pitfall of evaluation design, where budget or other considerations dictate the design of the study and the MDE simply

[41] Note that this can occur despite the best efforts of evaluators to caution against over-generalization—the message that the mechanism may operate even though the particular intervention failed is a difficult one to communicate to policymakers.

follows from that.[42] Underpowered studies cannot support decisions about whether to pursue a program, and the mere fact of reporting the prospective power calculation in the postmortem evaluation report does little to repair this. Even when sample sizes and MDEs are dictated by non-study constraints, evaluation results are likely to better support policy decisions if they are contextualized relative to pre-specified threshold or other analyses of what effects would be programmatically meaningful.

Second, Barnow and Greenberg discuss at length the representativeness (or lack thereof) of the populations included in demonstration studies. A particular challenge is the reliance on volunteers for sample recruitment. This is a necessity in many demonstrations, particularly those involving changes to programs that are legal entitlements (as in many of SSA's demonstrations). Nevertheless, those who step forward to participate in a trial are likely those who see the largest potential benefits from the program being tested, greatly limiting our ability to generalize to the wider population. In other contexts, this has been called "randomization bias" (Heckman 1992; Malani 2006). I view this as a very serious problem and see two potential ways of dealing with it. First, sometimes redefining a study as a mechanism study can avoid the problem—if the goal of the study is merely to test whether a mechanism operates, perhaps it is enough to establish that it operates in *some* subpopulation. Second, we might consider varying the incentive to participate in the trial across sites or subpopulations and using this variation to test the magnitude of randomization bias, which will tend to decline as the incentive to participate grows. This is analogous to DiNardo et al.'s (2021) proposal for avoiding survey nonresponse bias.

[42] For example, the POD evaluation design report (Wittenburg et al. 2018) discusses a target of 9,000 participants as following primarily from logistical and budget concerns, then calculates MDEs based on this sample size. These MDEs are characterized as "relatively small impacts," but there is no formal or informal analysis to justify these MDEs as related to thresholds for program success.

Chapter 2

Comment

Jack Smalligan
The Urban Institute

Burt Barnow and David Greenberg (in "Design of Social Security Administration Demonstration Evaluations") have written a very impressive and thorough discussion of some of the past demonstrations conducted by the Social Security Administration (SSA), the evaluation methodologies SSA has used, and the evaluation techniques SSA should consider for future demonstrations. Their chapter reviews 16 SSA evaluations, including 12 using experimental assignment designs.

Barnow and Greenberg identify several ways in which SSA evaluations are unique from evaluations of other social programs. First, the focus for SSA's demonstrations are individuals receiving or potentially receiving Social Security Disability Insurance (SSDI), Supplemental Security Income (SSI), or both benefits. This focus has the advantage of SSA evaluations often being able to use SSA administrative data, but it also introduces limitations that I will discuss below. Second, participation in SSA evaluations is voluntary. In contrast, evaluations in the Unemployment Insurance (UI) program, Temporary Assistance for Needy Families program (TANF), and Supplemental Nutrition Assistance Program (SNAP) are mandatory and the high turnover rates in the programs broaden the target audience.

Barnow and Greenberg discuss a range of evaluation techniques that SSA can explore for future demonstrations, including alternative experimental designs and clustered and adaptive designs. They also identify some seldom estimated impacts that SSA could include in future demonstrations. Regarding entry or deterrent effects, where an intervention may encourage or discourage participation in SSDI or SSI, they recognize that these effects are hard for SSA to measure given the target population of individuals already participating or potentially participating in its programs. However, they conclude, "The modesty of these estimates…suggests that entry and deterrent effects need not be considered a major issue in SSA evaluations." This conclusion I will revisit in the discussion below.

To put Barnow and Greenberg's conclusions in a broader context, I'm going to consider the design framework for SSA demonstrations and focus on how we re-envision the federal government's overall demonstration research agenda for people with disabilities. In short, the framework for SSA's demonstrations should be broadened, in terms of both the target population and the types of program features that are evaluated.

First, the programmatic focus for federally funded demonstrations should broaden. As Barnow and Greenberg discuss, the current unit of analysis for SSA's demonstrations is individuals receiving or potentially receiving SSDI, SSI, or both benefits. Congress should instead view this as national demonstration authority. Many

more Americans identify as having a disability compared with the subset of individuals participating in SSDI and SSI or seeking to participate in the programs. If Congress gave a broader charter, more demonstrations could test and evaluate interventions where programs intervene earlier with at-risk individuals who have no connection to SSDI or SSI.

The US Department of Labor's Office of Disability Employment Policy (ODEP) and SSA have made a start on a broader focus with the Retaining Employment and Talent after Injury/Illness Network (RETAIN) demonstration. RETAIN seeks to intervene with at-risk workers long before they have any connection with SSDI or SSI. ODEP is funding the intervention itself, and SSA is funding the evaluation—a complicated arrangement that enables ODEP to fund services for individuals with no connection to SSDI or SSI. Congress could expand SSA's Section 234 demonstration authority to fund evaluations for workers at risk of needing support from SSDI or SSI, allowing SSA to fund inventions that complement what ODEP is funding.

A variety of disability experts have proposed demonstration projects that could be tested using this broader authority. Christian, Wickizer, and Burton (2016) propose the "establishment of a community-focused Health & Work Service…dedicated to responding rapidly to new health-related work absence" (1). Stapleton, Ben-Shalom, and Mann (2016) propose "the development, testing, and adoption of a nationwide system of integrated employment/eligibility services" (21).

Looking ahead, policymakers have a strong interest in expanding access to paid medical leave, in addition to parental and caregiving leave. More states have enacted comprehensive paid leave programs, and proposals for a national program are growing.

Although most workers who take medical leave return to their jobs quickly, research shows that some are at an increased risk of leaving the labor force and experiencing serious hardship. Although the ability to take time off with pay is critical for these workers, return-to-work services could provide an opportunity to improve their health and employment outcomes. Should Congress enact a national paid leave program, the agency Congress directs to administer the program should be given authority to test and evaluate how to deliver those services (see Smalligan and Boyens 2020).

Second, in terms of SSA-specific demonstrations, we need to examine SSA's own internal eligibility determination process. Researchers should design process evaluations that are not evaluating a new intervention but are evaluating SSA's own internal disability eligibility determination processes.

For many years SSA's determination process faced backlogs, with eligibility determinations taking some workers one to two years. Research by Autor and colleagues (2015) shows that these delayed decisions lead to a decay in the work capacity of denied applicants. In other words, SSA's own eligibility determination process functioned essentially as an intervention with adverse employment outcomes for denied applicants.

SSA's existing Section 234 demonstration authority is explicitly linked to return to work. Congress needs to broaden the 234 authority so that SSA can redesign the process to function better and evaluate those efforts. In doing so, SSA could learn whether we can invest more in making better decisions, at an earlier stage. Earlier I summarized Barnow and Greenberg's discussion of possible entry and deterrent effects from interventions. SSA's arduous determination process may create a deterrent to applying for benefits, especially for people with barriers. For example, the closure of SSA's field offices during the COVID pandemic resulted in a substantial drop in SSI applications, suggesting low-income individuals are especially disadvantaged by obstacles to interacting with SSA.

The reconsideration stage of SSA's determination process could be used to test multiple approaches to an enhanced determination process. The goal of an enhanced second-level review would be to achieve better decisions earlier than are achieved today. The additional time spent developing a case at the state disability determination service level might be particularly important for applicants with low incomes and no health insurance. These claimants might have little or no medical evidence of record and a more difficult time presenting their case during an initial and second-level review and might otherwise need to wait for a decision at the hearing level.[43] SSA Commissioner Jo Anne Barnhart (2001–2007) began testing an effort to enhance the second-level review, but the effort was terminated by Commissioner Michael Astrue (2007–2013) before the results could be fully evaluated (Smalligan and Boyens 2019).

Congress should expand the Section 234 demonstration authority to permit testing and evaluating an enhanced disability determination process. This would be a substantial expansion of SSA's demonstration authority and requires SSA to consider creative evaluation techniques. Under this expanded authority, Section 234 would provide funding for the marginal additional cost of an enhanced determination process as well as the usual cost of a rigorous evaluation. SSA's administrative budget is always constrained and providing SSA the ability to test and evaluate new approaches without cutting back other activities would facilitate experimentation. This is a second area that requires Congress to redesign the existing SSA demonstration authority.

---

[43] The *hearing level* is the level following reconsideration in the administrative review process. The hearing is a de novo procedure at which the claimant, the claimant's representative, or both may appear in person, submit new evidence, examine the evidence used in making the determination under review, give testimony, and present and question witnesses. The hearing is on the record but is informal and nonadversarial (SSA 2020b, Glossary).

Chapter 3

Improving the Use of Demonstrations

Robert R. Weathers II
Social Security Administration[1]
Austin Nichols
Abt Associates

The Social Security Administration (SSA) has made substantial investments in planning and conducting demonstration projects. In 2008, the Government Accountability Office (GAO) found that between 1996 and 2008, SSA spent $155 million on demonstrations that "yielded limited information on the impacts of the program and policy changes they were testing" (GAO 2008, 1). The 2008 GAO report recognized that SSA had taken steps to improve its demonstrations, and SSA responded to the report by developing written policies and procedures for managing and operating them consistent with research practices and internal control standards in the federal government. SSA's response to the GAO report established a solid foundation for improving the evidence drawn from demonstrations on the efficacy and effectiveness of program and policy changes.

SSA has completed five demonstrations since 2008: the Mental Health Treatment Study (MHTS), the Youth Transition Demonstration (YTD), Accelerated Benefits (AB) demonstration, the Benefit Offset Pilot Demonstration (BOPD), and the Benefit Offset National Demonstration (BOND), all described at length in this volume's Appendix. Each one has provided rigorous evidence on the effects of the program and policy changes that were tested. SSA spent more than $245 million to complete these five demonstrations, and GAO considered all of them to be either "strong" or "reasonable" relative to professional research standards. SSA has a public-facing website that contains information on these demonstrations. It has published findings on them in peer-reviewed professional journals, highlighted their findings in its annual performance report and the annual report on the Supplemental Security Income (SSI) program, and produced annual reports to Congress that document progress on the demonstrations and the key findings from them.

SSA has made meaningful progress on planning and conducting demonstrations and continues to do so. However, there are several opportunities to further improve the return on these investments. This chapter identifies specific areas where SSA could make improvements to its demonstrations and broaden the evidence base used to make important program and policy decisions. We organize the specific areas under four headings: (1) tailoring the design of demonstrations to improve the use of results

[1] The views expressed in this chapter are those of the authors and do not necessarily represent the views of the Social Security Administration or the US federal government.

(including several cost-benefit considerations); (2) identifying and acquiring the data needed for evaluation; (3) expanding the dissemination of findings to stakeholders; and (4) broadening the use of data from the demonstrations to inform program and policy development. In addition to identifying specific areas for improvement, we provide examples from other research and demonstration efforts on how similar efforts have increased the evidence necessary to inform programs and policies. We also note areas where SSA has led the field in developing demonstration best practices, which implies it could continue to be a leader in improving the use of demonstrations.

TAILORING THE DESIGN OF DEMONSTRATIONS TO IMPROVE THE USE OF RESULTS

A demonstration design report provides a blueprint for the implementation of (1) the new or changed policy, program, service, support, or procedure (the "intervention") at the center of the demonstration, and (2) the evaluation of that intervention. A good design report describes the intervention to be evaluated, the intended effects of the intervention, the evaluation methodology, the structure of a cost-benefit analysis, the data sources, the implementation plan, and the dissemination plan. Failure to establish a sound demonstration design generally results in a failed demonstration.

We identify several aspects of a demonstration that need to be considered for inclusion in the design report to maximize the project's value to stakeholders. This section begins by considering the specification of the intervention. It describes instances where an intervention that provides information on a range of options can be more informative than an intervention focused on a specific option. We acknowledge that an intervention that informs a range of options might not be a practical policy option due to factors such as the incentives to participate and costs. However, such an intervention can have the advantage of reducing the set of effective options. For example, if a generous program proves to be ineffective, then less generous variants are unlikely to be effective.

We then consider ways to make greater use of theoretical and logic models to provide policymakers with more information on the potential efficacy and effectiveness of an intervention. Next, we examine how evaluation methods should be tailored to the intervention, as well as the need for evidence providing the rationale for the demonstration in the first place. We conclude this section by describing important components of a cost-benefit analysis that will allow stakeholders to assess the potential value of an intervention relative to its costs.

Designing Demonstrations That Inform a Broader Range of Policy Options

Some of SSA's demonstrations have focused on a narrow range of policy options, which can produce relatively limited information on the potential impact of alternative policies when compared to projects that inform a broader range of options. Most of

the focus is on SSA policies pertaining to Social Security Disability Insurance (SSDI) beneficiaries and SSI recipients, and often on their work outcomes, or rules related to work.

For example, BOND was designed to estimate the impact of changing the way that work behavior affects benefit payment amounts. The "benefit offset" refers to the change in rules associated with the benefit calculation. Under the current program rules, if a beneficiary performs work that is determined to be Substantial Gainful Activity (SGA), SSA suspends the entire benefit payment amount after a nine-month Trial Work Period and a three-month Grace Period. The complete loss in benefits is often referred to as the "cash cliff." Instead of suspending the entire benefit payment amount, BOND Stage 1 tested the impact of a more gradual $1 reduction in benefit payments for every $2 in earnings above an annual amount corresponding to the SGA level. Thus, BOND Stage 1 tested this specific benefit offset policy against the current policy.

However, there are a number of variations to a benefit offset policy that might be of interest to policymakers, and that are likely to have different impacts on work activity, benefit payments, and the finances of SSA's disability programs. For example, an even more gradual $1 reduction in benefits for every $4 in earnings above the SGA level might be of interest to policymakers because it might provide a relatively greater inducement to work.[2] Or, as is the case with the Promoting Opportunity Demonstration (POD), starting the benefit offset at a lower earnings amount than the SGA level might be of interest to policymakers because it may be more likely to result in program savings. Variations in the benefit rules provide different incentives for SSDI beneficiaries to work. Other policies or programs could also change the effectiveness of a specific offset, greatly expanding the range of policy options to consider.

More generally, a focus of SSA demonstrations, as well as changes to the work incentives within the SSI and SSDI programs, is the assumption that SSDI beneficiaries and SSI recipients with a capacity to work do not do so because of the loss in benefits associated with work activity.[3] The loss in benefits associated with work activity is implicitly a tax on earnings. This "tax" lowers the relative price of non-work activity, which could encourage beneficiaries and recipients to reduce time

---

[2] SSA's SSDI demonstration project authority described in Section 234 of the Social Security Act specifically identifies potential alternatives such as "implementing sliding scale benefit offsets using variations in—(i) the amount of the offset as a proportion of earned income; (ii) the duration of the offset period; and (iii) the method of determining the amount of income earned by such individuals." SSA has used the SSI research authority in Section 1110 of the Social Security Act to test a $1 reduction in SSI payments for every $4 in earned income as part of YTD.

[3] Data from the National Beneficiary Survey show that 91.7 of all SSI recipients and SSDI beneficiaries identify a physical or mental health condition as the primary reason for not working (SSA 2018a; prepared by Emily Roessel, Office of Research, Demonstration, and Employment Support).

spent working. The implicit tax on earnings Autor and Duggan (2007) refer to as the "substitution effect" channel for discouraging work activity, by which they mean the tax leads beneficiaries and recipients to substitute from time spent at work with time spent at their leisure. Alternatively, access to the SSDI cash benefit and Medicare can also discourage work activity. The availability of cash benefits and Medicare is referred to as the "income effect" channel for discouraging work, as the income from them subsidizes leisure activities and can be the relatively more important factor that reduces work activity among beneficiaries and recipients. If the income effect channel is the primary factor affecting work activity among beneficiaries, then SSA's demonstrations that focus on the substitution effect will be relatively ineffective in encouraging work.[4] The magnitudes of these two types of beneficiary and recipient responses are important for understanding the implications of proposed work incentives policy changes (Gelber, Moore, and Strand 2017).

SSA's demonstrations to date have tested the importance of the substitution effect channel in a limited and incremental manner. The projects have focused on a $1 for $2 offset as opposed to other sliding scale offsets (e.g., a $1 reduction for every $4 earned), and on a limited range of the amount of earnings allowed before the benefit offset begins. One reason for the focus on a $1 for $2 benefit offset policy is because the law explicitly required testing such an offset. In addition, limited resources were available for conducting a wider range of benefit offset demonstration projects.

A consequence of the focus on a $1 for $2 benefit offset is that we do not know what effects other kinds of policy innovations might produce, and we do not know what the limits are on the effects that could be produced by altering incentives tied to the substitution effect channel. We also do not understand the "response surface," or how effects on beneficiaries depend on the detailed parameters of policies. That is, the existing crop of demonstrations addresses a specific subset of questions specified in the law. If new legislation provides the agency with more discretion, new demonstrations could be used to greatly widen the range of questions to which we have answers, both by answering more questions in each demonstration and by increasing the scope of questions answered to explore the limits on possible effects of a type of intervention.

As an example of testing the limits on possible effects, a demonstration could allow beneficiaries to maintain their entire benefit amount and Medicare no matter how much they earn. This intervention would completely eliminate the implicit tax on earnings due to program rules. The actual policy tested by such a demonstration might not be viable, due to the potential effects on program participation and program costs. However, the demonstration could provide a plausible upper-bound estimate of the

[4] There is a long history of attempts to measure the impact of disability benefits on lowering work activity, though it is not clear whether this is due merely to the increased income due to receipt of benefits, rather than any disincentive to the effective taxation of labor income (Bound 1989; Bound and Burkhauser 1999; von Wachter, Song, and Manchester 2011; Maestas, Mullen, and Strand 2013; French and Song 2014).

impact of different benefit offset policy options (and other policies that change the economic benefit of work) on the number of beneficiaries who perform work considered SGA. Eliminating benefit reductions removes the substitution effect channel that discourages work and provides evidence on the relative magnitude of the substitution effect channel (i.e., the tax on work activity is the primary work disincentive) versus the income effect channel (i.e., the benefit amount and access to Medicare is the primary work disincentive).

To better understand the way SSDI beneficiaries and SSI recipients respond to any possible variations in policy, a demonstration can use a multi-armed or factorial experimental evaluation design (see Chapter 2 in this volume). Doing so introduces variation in the benefit calculation rules, with several varying amounts, or additional interventions such as services that might make new benefit calculation rules more or less effective at increasing work. A good example of a multi-arm design that SSA successfully deployed is BOND, in which Stage 1 tested just the new policy for annual earnings, whereas Stage 2 tested a pair of policy changes for volunteers only. The first experimental arm got the same intervention as Stage 1, and the control arm also got Stage 1's business-as-usual condition. A second experimental arm in Stage 2 received both the $1 for $2 offset and more proactive counseling regarding benefits and work, called "enhanced work incentives counseling," at an increased cost. Stage 2 did not find any evidence that the extra counseling increased earnings, so evidently the combination of proactive counseling and the offset is inferior to the simple offset on cost-benefit grounds. This is information that would not have been knowable without the additional arm of the evaluation.

Broadening the Use of Theoretical Models

Specifying a theoretical model is useful for describing the potential effects from a demonstration, specifically what response to an intervention to expect. For example, for the BOPD and BOND, a simple static labor supply model illustrates the potential effects of a change to the benefit offset policy for current beneficiaries on earnings amounts and benefit amounts (Weathers and Hemmeter 2011). A life-cycle model is useful for describing broader behavioral effects of an intervention, such as how it might lead to "entry effects" (see Chapter 2)—that is, it might encourage some individuals to apply for benefits sooner than they otherwise would under current program rules (Benítez-Silva, Buchinsky, and Rust 2010).

One opportunity to extract more information from demonstrations is to use the results to estimate and validate a well-specified theoretical model. For example, Todd and Wolpin (2006) use the results from a social experiment on a school subsidy program to estimate and validate a model of parental decisions about fertility and a child's educational attainment. They then use the model to analyze different policy proposals and provide information on the likely impacts of alternative policies on fertility and on educational attainment.

There are a number of different theoretical models that could be estimated and validated to inform changes to SSA policy. For example, SSA can use results from benefit offset demonstrations to estimate and validate a life-cycle model similar to the one specified by Benítez-Silva, Buchinsky, and Rust (2010). SSA can then use the model to simulate the effects of different types of benefit offset proposals on outcomes such as entry into the program, work behavior, benefit amounts, and potential program costs or savings. For example, the ongoing Promoting Opportunity Demonstration (POD) is using a model to simulate different policy effects at a national level, but as far as we know, is not simulating entry effects. Another example is related to the timing and the amount of payments to service providers, referred to as "Employment Networks," under the Ticket to Work program. If we have information on how specific changes to the Ticket to Work payment structure influenced Employment Network behavior, we could estimate and validate a theoretical model on the relationship between the Ticket to Work payment structure and the behavior of Employment Networks. We could then use the model to simulate how other possible changes to the payment structure would change the size and composition of the number of SSDI beneficiaries and SSI recipients served and their employment outcomes.

The amount of confidence in results of model-based simulations depends on the information used to estimate and validate the model. For example, using results from several demonstrations that tested different benefit offset policy options could result in more reliable model estimates. Specifically, testing more generous benefit offsets than have been tested to date would require extrapolation with no empirical support; but as described in Chapter 1, the "Ultimate Demonstration" (see Gubits et al. 2019) would provide information so that results for almost any offset policy would have empirical support. Therefore, our suggestion that SSA design demonstrations that test a wider range of policy options would be a beneficial input to such a model-based approach.

A well-constructed theoretical model is also useful for assessing how program interactions can affect responses to a program or policy. There are two aspects of BOND where a theoretical model is particularly useful. One is the interaction between SSDI and SSI. In 2018, approximately 14 percent of SSDI beneficiaries received benefits from both programs. Because the SSDI benefit amount is treated as unearned income when determining the SSI payment amount, the drop in SSDI benefits that occurs when a beneficiary performs SGA can result in an increase in the SSI payment amount, which lessens the implicit tax on work activity under the existing program rules. Thus, BOND might not reduce the disincentive to work among SSDI beneficiaries who also receive an SSI payment.

Another potentially important interaction occurs for SSDI beneficiaries whose income levels make them eligible for Medicaid benefits or health insurance subsidies through a state Medicaid buy-in program or through the Affordable Care Act. Work activity could affect eligibility for such benefits, and subsequently affect the financial incentive for beneficiaries to participate in SGA under BOND. While the BOND

evaluation considered such effects, the lack of individual-level data on participation in Medicaid buy-in made disentangling the effect difficult. Indeed, this was a limitation of BOND.

Establishing a well-specified theoretical model as part of the demonstration design can expand the information on a policy or program drawn from a demonstration. We identified at least three advantages of using such models. First, they can be used along with data drawn from a demonstration to produce simulated responses to alternative policy or program changes. Second, they can provide information on potential program or policy changes that could not be adequately measured from a demonstration, such as the potential for a policy or program change to encourage entry into the program. Third, they can identify important interactions with other programs that could limit the potential effect of a program or policy. Theoretical models can inform the demonstration's intervention design and then, thereafter, can provide information on implications of the evaluation's findings.

Broadening the Use of Logic Models to Specify Detailed, and Falsifiable, Goals

A logic model lays out an intervention's inputs, activities, outputs, and outcomes. "Inputs" include the funding or legislative authority, for example, that makes the intervention possible. Inputs included the funding, staff, and mechanisms to implement the intervention. The logic model then identifies the activities (such as counseling services) and "outputs" (such as that participants receive counseling). Next the logic model identifies the outcomes, including both short- and long-term ones, that the intervention aims to influence. A logic model can also show external factors that moderate impacts or could interfere with the linkage of inputs to outputs and outcomes. The logic model is helpful for understanding the intervention at the center of a demonstration, and how to measure that intervention's success. Ultimately, the logic model is a high-level summary of the assumptions about how the intervention is expected to operate.

SSA has specified logic models prior to enrolling participants into an intervention, and those models provide a clear picture of the expected relationships on which the evaluation focuses. The SSA demonstrations with well-specified logic models include the AB demonstration (see Weathers et al. 2010, Chart 1), BOND (see Stapleton et al. 2010, Exhibit 2.5), the Promoting Readiness of Minors in Supplemental Security Income (PROMISE) demonstration (see Fraker, Carter, et al. 2014, Figure I.1), and YTD (see Rangarajan et al. 2009, Figure I.1). Yet they are often missing links that could have been useful to identify early measurement supporting program improvement.

As a specific example, consider YTD. As Hemmeter (2014) summarized:

> Most of the types of services provided at YTD projects were those recommended by the National Collaborative on Workforce Disability for Youth, although some were drawn from "best

> practices" of other interventions for youths with disabilities. The YTD project's core interventions addressed the barriers youths face in their transition from school to work. Chart 1 ["YTD Design Objectives"] depicts the barriers and the YTD intervention components, along with the transition environment and key project outcomes.

In the figure, barriers (such as low expectations about work and self-sufficiency or financial disincentives to work), YTD intervention components, and factors affecting transition (such as schools or community-based service providers) are listed in three separate boxes that each point at a central oval marked "Transition Efforts by Youth," which then points to short-term and longer-term outcomes. But

> each of the YTD sites offered services to break down…barriers to varying degrees…[e.g.,] empowerment training to help participating youths learn to make their own choices (as opposed to having a parent or guardian choose for them)…; working with the families to break down misunderstandings about program rules; encouraging the families to participate in planning for the youths' self-sufficiency…; and providing case management to coordinate health and other social services.

Evidently, there are a number of hypothesized links in the causal chain that are not spelled out in the three boxes pointing at a central oval in the logic model. For example, presumably "encouraging" families to plan for their youth's self-sufficiency is intended to shift the "low expectations" identified as a barrier in the logic model. But Wood and Goetz Engler (Chapter 9 in this volume) report that "more than 80 percent of enrollees reported that they expected to work at least part-time in the future" and "a more generous $1 for $4 benefit offset in the earned income exclusion and an extension of the student earned income exclusion…encouraged participants to enroll." That is, YTD participants may have come in with high expectations about work and independence, and counseling could actually have shifted them in the wrong direction, or not at all. Pre- and post-counseling measures of expectations would have provided direct feedback on this link in the chain.

As Martinez et al. (2010) note, in some YTD sites, "the provision of direct employment-related services such as job development was not a primary focus of the program intervention, yet independence and self-sufficiency were cited as primary goals. Clearly defined pathways that would suggest that the proposed services could directly lead to self-sufficiency were not evident." If each site could not say how services might lead to self-sufficiency, and measure changes that might be expected one day to lead to increases in paid employment and income (or even immediate changes in attitudes and expectations), the site's logic model is evidently missing some testable hypotheses.

An extension of the basic logic model can provide an opportunity to connect impacts on short-term or long-term outcomes, or even near-term changes in attitudes or participation, to intervention components. The extension is referred to as the "falsifiable logic model" (FLM) and differs from logic models specified in prior SSA demonstrations because it includes "the requirement that an expanded logic model specify detailed—and falsifiable—goals for one of the components of a conventional logic model—intermediate outcomes that must be realized by members of the treatment group in order for the program to succeed" (Epstein and Klerman 2012, 380). Examples of such intermediate outcomes could include achieving a pre-specified target for the use of a specific intervention (not all who enroll choose to use a specific intervention), achieving a pre-specified goal of successful completion of the components of an intervention, or other intermediate outcomes that the logic model identifies as a key, or a combination.

There are several advantages to specifying detailed goals within the FLM. First, pre-specifying key intermediate outcomes establishes metrics that could provide early information on the potential effectiveness of the intervention. If it fails to achieve the intermediate outcomes that are intended to lead to ultimate outcomes, then we would expect smaller ultimate impacts of the intervention on ultimate outcomes. For example, the PROMISE demonstration focused on services delivered in the intermediate report since its logic model specified those services were needed for the intervention to have longer-term impact (Mamun et al. 2019). Or, if there are immediate positive findings, the intervention might be judged favorably before long-term outcomes are collected, e.g., the denied applicant mailer study de-prioritized longer-term analysis because appeals fell in the near term (GSA 2019).

Second, it provides metrics that point to areas for future changes to an intervention to improve its efficacy. Third, if the FLM is part of a pilot project, then it might provide the additional information necessary to make decisions about whether, as well as how, to proceed with more rigorous impact evaluation. Fourth, Epstein and Klerman (2012) identify the FLM as a mechanism for "truth-telling" during the course of the demonstration. That is, specifying intermediate outcomes as part of an FLM could reduce the tendency to initially oversell an intervention to promote its implementation, and it could reduce the tendency to understate expected effects of an intervention at the evaluation stage to claim success. A clear example is in recruitment: if fewer than six in a hundred eligible participants agree to use any offered services, the impacts on those six would have to be enormous to produce discernable population-level impacts. Solicited participants signing up for a demonstration, and then using services offered, are common early links in a chain of events that are indicative of upper bounds on long-term success.

Implementation of a FLM might have been useful for past demonstrations, as well as for evaluations of ongoing national programs. For example, the AB demonstration included the Progressive Goal Attainment Program (PGAP), which consisted of 10 modules that were designed to progressively change a participant's behavior and

increase the likelihood of their return to work (Michalopoulos et al. 2011). The AB logic model did not specify metrics for intermediate outcomes for PGAP in terms of the numbers who would participate and the number of modules that participants would need to complete for PGAP to contribute to employment outcomes. At the end of the evaluation, the final report showed that 36 percent of those eligible for PGAP used it, only about one-sixth of that group completed all 10 modules, and half completed at least four modules. An FLM that identified four modules as a sufficient amount to have an effect on employment would have strengthened the findings. That is, if the demonstration could have identified up front the level of fidelity likely to produce discernable impacts, early results on the fraction of participants attaining that level of fidelity to planned services would have helped to set expectations about the upper bound on eventual long-term impacts observable.

It seems likely to us that an FLM might also have improved the information drawn from evaluations of other demonstrations—such as Project NetWork, and YTD as described above. In Project NetWork, only 4.5 percent of those solicited signed up, but "60 percent of participants completed assessment and employment planning and 45 percent received purchased employment-related services," as detailed in Chapter 9. That is, the offer of service might result in fewer than three in a hundred getting assessments, and if "there were delays in obtaining the initial assessments of participants, [they could well disengage] during these waiting periods." Had the demonstration tracked these milestones on the path to impact, they might have tweaked processes to improve adherence. These are the low-hanging fruit, easily gleaned via process reports. Getting inside the delivery of services, and measuring the quality of services delivered, should also be part of the FLM. A useful analogy might be education: if students do not attend class, then the quality of the teaching does not matter. If we can track attendance easily, we should; that measuring the quality of teaching is a harder nut to crack does not mean we do not try.

In the MHTS, researchers (Frey et al. 2011) used fidelity measurement to assess site-level service delivery of the manualized intervention Individual Placement and Support (IPS) and to improve implementation in progress. The researchers conducted annual site visits and rated them on adherence to IPS standards related to staffing, organization, and service requirements. The Supported Employment Demonstration (SED) also uses the IPS intervention, which has a rich literature supporting its effects in various populations and a fidelity scale that can be used to identify shortfalls early in the logic model. In a sense MHTS and SED illustrate both the promise and pitfalls of the FLM, since well-measured service delivery outcomes can be used to course correct and improve interventions early in the demonstration, but these can alter the intervention as the demonstration is ongoing (complicating interpretation of any evaluation). This approach could, for example, prune sites where impacts are unlikely due to faulty implementation, breaking the early links in the causal chain, but that would change the nature of the impact being evaluated. The purposive site selection in MHTS may play a similar role, also changing the nature of the impact being

evaluated, from the impact of IPS on those with schizophrenia or affective disorder and no employment to the impact in just those kinds of sites chosen for the ability to deliver IPS with high fidelity.

We believe that there are opportunities for SSA to use FLMs prospectively to help assess interventions that are good candidates for demonstrations. For example, SSA is currently conducting a study on exits from disability assistance. This study may identify potential interventions that might promote economic self-sufficiency for SSDI beneficiaries and SSI recipients who are no longer eligible for disability benefits because they had a medical review that indicates they are capable of performing SGA. The results of this study will be useful if an FLM can be specified with multiple links in service delivery and outcomes that can support both early detection of more and less promising models and eventual rigorous impact evaluation. Similarly, SSA is assessing early intervention efforts that might help people with disabilities obtain and maintain employment and reduce the likelihood of applying for disability benefits. SSA can use an FLM for potential early intervention pilot projects, to choose the most promising interventions based on intermediate outcomes and recruitment, well in advance of any one of those interventions being part of a larger demonstration. There are plans to pursue such projects in the near future, so this is a ripe opportunity to pilot new FLM approaches.

Another opportunity for SSA to incorporate FLMs is through its new Interventional Cooperative Agreement Program (ICAP). Its establishment will allow SSA to collaborate with states, private foundations, and other entities that have the interest and ability to identify, operate, and partially fund interventional research. The research and interventions under ICAP will focus on: examining the structural barriers in the labor market, including for racial, ethnic, or other underserved communities in addition to people with disabilities, that increase the likelihood of people receiving or applying for SSDI or SSI benefits; promoting self-sufficiency by helping people enter, stay in, or return to the labor force, including children and youth; coordinating the planning between private and human services agencies to improve the administration and effectiveness of the SSDI, SSI, and related programs; assisting claimants in underserved communities apply for or appeal determinations or decisions on claims for SSDI and SSI benefits; and conducting outreach to children with disabilities who are potentially eligible to receive SSI as well as their parents and guardians. The awards will be tiered, with funding eligibility and the level of funding based on the level of evidence that currently exists for the proposed intervention (i.e., feasibility studies with no causal evidence would be eligible for smaller awards than studies scaling up or otherwise implementing interventions that qualify as “effective” according to statistical and evaluation criteria). SSA could use FLMs as a mechanism for assessing the readiness of pilot projects for rigorous evaluation, and depending on the number of applications, a random assignment of FLM to pilots might even be feasible in this setting.

Documenting the Tradeoffs When Defining the Scope of a Demonstration

In addition to using theoretical and logic models to inform evidence-generating efforts, SSA also must consider various practicalities when planning its demonstrations. There are sharp limits on what SSA is authorized to test, and those limits are slated to become much sharper when its current demonstration authority expires next year. SSA as an agency also has to be sensitive to the priorities of numerous stakeholders who fund it, audit it, or comment on its rulemaking. Beyond these legislative and political barriers, there are important statistical and funding constraints on what can go into a demonstration. For example, in some instances a demonstration might not be practical for producing the evidence on a specific outcome that is important to policymakers. This could be because the effect size that is important to policymakers is small (too small to expect an evaluation to detect it, given the sample size). Another reason is that some effects of interest to policymakers cannot be easily evaluated because a demonstration setting cannot adequately approximate the conditions of an ongoing national program, or because the time frame for measuring relevant effects is too long. In other cases, the intervention might consist of several components, and resource limitations might make it too difficult to unpack the effect of each component. Below, we illustrate how to focus a demonstration to maximize the evidence it produces, subject to various practical constraints.

One example of an effect that is important to policymakers and that is difficult to estimate from a demonstration is the effect on the number and composition of program participants (i.e., entry effects; see also Chapter 2).[5] A change to the program that expands the benefits available to program participants could induce those who are potentially eligible for the program to enter the program, thus changing the number and composition of its participants. The way this effect unfolds can depend on how credibly permanent the policy change is, and how information about the change filters out to *potential* participants. For instance, in the SSDI program, adding a benefit offset, eliminating the 24-month Medicare waiting period, or adding other benefits might induce those with severe health impairments to exit the labor force and to enter SSDI. A change in the number or composition or both of program participants could change the average benefit payment amount, the average duration of program participation, and the number of participants, and collectively influence the costs of the program.

Using a demonstration to estimate entry effects would require reaching those individuals who are not participating in the intervention, meaning the sampling frame for a demonstration would need to be the entire population. In the entire population, however, most potential participants have very low chances of ever participating, even given large changes in incentives. As a result, the average impacts are a mix of "zeros" (negligibly small changes in chances of entering the program) and larger impacts,

[5] Section 302 of the Ticket Act refers to program entry effects as "induced entry," and required SSA to conduct a benefit offset demonstration project that includes an evaluation of "the effects, if any, of induced entry into the project and reduced exit from the project."

which implies an extremely large random sample of the entire population would have to be randomized in order to detect impacts. Some narrowing of the scope of the demonstration is required to make progress.

BOND is a good example of how to narrow a demonstration to provide useful information to policymakers. The Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act) specified a benefit offset demonstration project of sufficient size and scope to estimate "the effects, if any, of induced entry into the project and reduced exit from the project." SSA actuaries have traditionally assumed that the costs associated with entry effects are larger than the assumed savings due to increased work activity (McLaughlin 1994). During the design phase of the benefit offset demonstration, SSA faced an important decision: whether a demonstration project would be a practical and reliable way to estimate entry effects under an ongoing national benefit offset policy and should be conducted, or whether SSA should pursue a narrower demonstration focused on the effect of a benefit offset on those who enter the program under the existing rules.

SSA conducted a considerable amount of research and analysis to inform a decision on the type of benefit offset demonstration project it should pursue. A team of experts reviewed SSA's work, conducted their own analysis, and summarized their recommendations and conclusions in a report (Tuma 2001). That work identified how SSA should think about the target population for each demonstration. A demonstration project designed to estimate entry effects would need to target the population not participating in the program. Without a practical way of limiting the pool of nonparticipants to those who are potentially eligible for SSDI, the demonstration would need to target a sample from the US population. The expert panel noted that such a target population poses challenges because nonparticipants are much more numerous than current participants, and because current participants are more numerous than the number of induced entrants that would contribute to substantial program effects (costs). Thus, a demonstration project would need to be very large—on the order of nine million people—to detect meaningful program effects of the magnitude that would likely arise or be policy relevant or both (Tuma 2001).

The experts also confirmed that an experimental design to measure entry effects for the demonstration's evaluation would need to randomize by geographical units rather than by individuals in the target population. Randomization at the individual level would require a method of informing millions of individuals about the new program in a way that would approximate the dissemination of information in an ongoing national program. The consultants agreed that contacting and explaining the new features of the program to individuals in the general population (with no prior contact with SSA, in many cases) randomly selected to participate in the program was impractical. The expert panel concluded that randomization at the county level, where information could be disseminated to those in the county by SSA field offices and by other means in a way that reasonably approximated dissemination in an ongoing program, might be a feasible design. However, the size of the demonstration would

make it potentially very expensive to administer, and other potential threats such as county-to-county migration during the demonstration were determined to pose risks to the demonstration related to the inability to reliably evaluate effects and generalize to the national population in the presence of individuals moving from one treatment status to another (known as "crossover"), or interacting with others and learning about other treatment conditions (known as "interference"). Because of those issues of cost and implementation challenges, a large-scale, national experimental evaluation of entry effects is unlikely in the future.

The alternative to a demonstration with an evaluation that directly estimates entry effects is a relatively narrower design that estimates the effects of changes to benefit rules on those who enter under the existing program rules. If the results from a benefit offset demonstration indicate that it is ineffective at increasing work activity and reducing benefit payments, then program entry effects may not be informative to the decision-making process as they would increase administrative and program costs.[6] If the benefit offset were effective at encouraging work activity and reducing benefit payments among existing beneficiaries, then SSA would pursue other research initiatives, such as the use of dynamic modeling of individual behavior, responses to hypothetical questions in a survey, or a stated preferences research design (Maestas, Mullen, and Zamarro 2010). SSA determined that a demonstration that estimates the work activity, benefit payments, and program costs was a more manageable and practical way to proceed; and that the demonstration would provide policymakers with the information needed to assess the effectiveness of changes to the benefit rules among current program participants. As it turns out, BOND showed increased benefits associated with both earnings increases and decreases that suggest entry effects could be important, even if potential entrants were in fact similar to current beneficiaries. POD suggests likely entry effects as well. Mamun et al. (2021) report that more than a third of POD participants who exited the demonstration (6 percent of treatment group members left) did so because program rules were more advantageous outside of POD. This strategic exit requires the same kind of calculation that potential entrants would face.

Looking Inside the Black Box

Experimental evaluation design provides an unbiased estimate of a true causal effect (rather than statistical noise or a confounded pattern), but it estimates the effect of the intervention which is a whole package of changes to the status quo, also typically known as the "treatment" in an evaluation setting. To measure the combined effect of a multi-faceted intervention, the standard A/B testing model divides evaluation

[6] The cost estimates produced by SSA's Office of the Actuary assumed that a benefit offset would be effective in increasing work activity and reducing benefit payments among current beneficiaries, but that the costs associated with program entry effects would be much larger and result in a net cost to the SSDI program.

participants (the "sample") into two groups. One gets a "treatment" that is the intervention's package of services or new policy parameters, and the other gets nothing new, or "business as usual." That treatment package is often referred to as a "black box" because one can see only the effects of the entire package on outcomes, not mechanisms by which the treatment produces the effects or which components matter in producing the effects. Such a black box could contain some elements that increase work and participant well-being and some other elements that decrease work and participant well-being, but the evaluator sees only the net effect.

In addition to a variety of quasi-experimental approaches to "looking inside the black box," there are two experimental approaches: using multiple distinct treatment arms as a collection of black boxes and using a factorial design, where program elements are systematically varied across multiple dimensions. We have good examples in the demonstrations of the multi-armed approach but lack good examples of factorial designs or quasi-experimental approaches.

Multiple Treatment Arm Approach

AB Demonstration. The AB demonstration provides a good example of the tradeoffs between the black box approach versus the multi-armed approach. That project included three randomly assigned groups of newly entitled SSDI beneficiaries who did not have health insurance coverage at intake: A group called "AB" received a health insurance package during the SSDI's 24-month Medicare waiting period; a group called "AB Plus" received the health insurance package plus additional rehabilitation and counseling services during the Medicare waiting period; and a control group received neither. The additional rehabilitation and counseling services that were available to the AB Plus group address the barriers that some newly entitled beneficiaries face as they attempt to return to work. Specifically, the AB Plus group had access to (1) medical care management along with the health insurance package to treat or stabilize their disabling health condition, (2) PGAP to encourage them to participate in activities that will eventually lead to work, and (3) employment and benefits counseling services to inform them of employment services and programs.

The design allowed SSA to estimate the effect of health insurance during the waiting period on outcomes of interest (AB group versus control group), as well as the effect of the additional package of services offered to the AB Plus group on those outcomes (AB Plus group versus AB group). The findings indicated that health insurance coverage alone was not sufficient to improve beneficiaries' labor market outcomes, but the addition of the AB Plus services was sufficient (Weathers and Bailey 2014).

The positive impact of the package of AB Plus services on labor market outcomes leads to questions on the relative importance of each of the three services. The effort to unpack the relative importance of each component relied on the choice among beneficiaries to use the additional services that were offered, providing relatively weak and limited evidence (Weathers and Bailey 2014). That approach suggested that those

who chose to use *employment and benefits counseling services* experienced substantially better labor market outcomes than those who chose not to use the service. That finding was consistent with those from other studies. Beneficiaries who chose to use *PGAP* or the *medical care management services* did not experience better labor market outcomes than those who did not use them.

The non-experimental approach used in that study provided evidence that is only suggestive. Understanding the relative effectiveness of the services, however, is very important for a cost-benefit analysis. This is because the medical care management services cost $1,312 per user, the PGAP services cost $1,734 per user, and the employment and benefits counseling cost $3,650 per service user (Michalopoulos et al. 2011). If any one of these services played no role in improving outcomes, then the elimination of them has the potential to reduce costs without a corresponding change in the outcomes of the program.

BOND. Stage 2 of BOND offers another example of a multi-armed approach in which added services play an important role. In Stage 2, two treatment groups received the benefit offset; one received standard work incentives counseling, but the other treatment group got enhanced work incentives counseling that resulted in a far higher proportion of them being told how the offset worked in detail with their potential labor earnings and benefits, given their family circumstances. That is, the enhanced counseling was the same kind of counseling offered in the standard work incentives counseling arm, but it was delivered to more people via proactive outreach.

Comparisons across the two treatment groups can illuminate the effects of extra counseling in the presence of the offset. The comparison of the standard work incentives counseling treatment group versus the control group (that also got standard work incentives counseling but no offset) illustrates the effect of the offset in this group of volunteers. The comparison of the enhanced work incentives counseling treatment group versus the control group illustrates the effect of the combined package of the offset and extra counseling.

Because those who were in the extra counseling group did not have discernably better outcomes, and the extra services were expensive, the cost-benefit analysis does not support the extra counseling as part of an implemented offset policy. Adding one more treatment group, that got enhanced work incentives counseling but no offset, would have made this a factorial design, which we discuss below.

We will just note here that there is likely a lot of interesting information in BOND Stage 2 that remains unexploited, beyond those simple comparisons of means (regression-adjusted or not), and the BOND study data seem ripe for further analysis. For example, understanding who volunteered for Stage 2 from the solicitation pool already provides information about the attitudes and expectations of those solicited. Participants could never be worse off in BOND, and stood to gain potentially thousands of dollars in net income in the situation where they successfully earned above the threshold, yet 19 out of 20 evidently did not value the opportunity to participate. There are a variety of subgroup comparisons of interest, and explorations

of intent to treat (ITT) versus treatment on the treated (TOT), that could be constructed for BOND. We suspect there are also long-term follow-up opportunities using the BOND data merged with other sources of information.

SED. The SED, which is still ongoing, used random assignment design to assign participants to one of two intervention groups (Full-Service or Basic-Service) or to a business-as-usual group. Over a three-year period, participants receive varying doses of services based on their random assignment. These services include systematic medication management, health care management and care-coordination services, and long-term employment services following the IPS model. Unlike the MHTS discussed in a prior section, SED will be able to say something about how effects vary with the intensity of services. Yet there are only two average levels to compare in this design, and the naturally occurring variation in each individual service cannot provide unbiased information about the value of each service, or any complementarities across services.

POD. The POD also included two treatment groups and a business-as-usual group. In POD, the first intervention group got an offset similar to BOND's, but starting at a lower earnings threshold, and could not lose entitlement to SSDI or Medicare no matter how much they worked. The second group got the same offset, but could lose entitlement to benefits if earnings proved too high for too long. The initial evaluation results from Mamun et al. (2021) indicate no differences in impact across the two intervention groups, and, for most of the findings, the two intervention groups are combined and compared to business as usual. The finding that adding or removing a protection from any potential loss of benefits has no impact is surprising, but since we can only compare the mean outcome across the two intervention groups, we learn little about the mechanism that generates this finding.

All of the demonstrations we discussed here, AB, BOND, SED, and POD, went beyond the standard A/B test to explore another dimension of variation in the intervention. Yet the multi-armed approach fundamentally still measures the effects of several black boxes, so we cannot extrapolate beyond the specific packages of interventions in each black box. Furthermore, exploiting the non-experimental variation in who used various program components provides less convincing evidence, hence conclusions about mechanisms that use that variation can only ever be "suggestive," as caveated above. The best quasi-experimental approaches to exploiting naturally occurring variation relies on strong assumptions to estimate the mechanisms via which an intervention produces effects (mediation) or factors that change its effects (moderation). VanderWeele (2011), and others in the same journal issue, explains some of the challenges to using principal stratification to isolate mediation. Similar analysis applies to the Analysis of Symmetrically Predicted Endogenous Subgroups method of Peck (2003, 2013). Questions about mediation or moderation frequently crop up immediately after the black box answer is known, because any intervention can be decomposed into constituent parts that could have quite different effects if combined differently. Importantly, the design of an experiment (see Chapter 2) only

addresses the specific research questions posed, so anticipating these questions is a matter of designing the research questions. If the design of the research questions dictates a multi-arm or similar design, presumably that is the design that researchers will settle on, if it is feasible.

Factorial Design Approach

Using factorial evaluation design to answer multiple questions, or to isolate the impact of components of an intervention on outcomes, provides a way to answer more nuanced research questions. That is, the most convincing way to estimate a mediation effect is to randomly assign multiple versions of assignment to treatment in such a way that there is random variation in take-up of different components of treatment. For simplicity, suppose there are only two components: benefit offset rates and counseling. The standard factorial design could assign, say, three types of benefit offset (e.g., business as usual, one-for-two above the SGA level, and no offset) and three types of counseling (e.g., business as usual, enhanced access to counseling, and proactive counseling, where the goal is that everyone gets the maximal level of counseling). This example can be represented then as a two-way tabulation of three rows and three columns that results in nine different treatment groups, one of which is business as usual in both dimensions (the control group). With many more levels of treatment, the number of groups expands exponentially, but the comparisons can remain reasonably simple looking at one dimension at a time.

Analyzing the resulting experimental data, we can compute how outcome vary with the intensity of intervention (called the dose-response function in most literature) for both dimensions independently (offset independent from counseling), and in this case, the advantage of the factorial design is that we get two experiments for the price of one. That is, we don't need to compute the required sample size to detect a policy-relevant impact for the first dimension of treatment using only the third of the sample that is business as usual in the second dimension, but instead can use the larger sample. However, this approach maximizes the power to detect differences in one dimension at a time. A factorial design also allows the evaluator to detect complementarities in treatments. For example, perhaps counseling has a greater effect on work outcomes in the presence of a more generous offset.

Policy Importance of Variation in Average Treatment Effects

SSA's demonstrations have focused on estimating "average treatment effects" of interventions, defined as the average effect of being offered a package of services or policies called the intervention or treatment. The prior section has proposed how demonstrations can be more useful at pulling apart the component pieces of an intervention for deliberate, rigorous evaluation. We suggest that demonstrations can be additionally useful by focusing on the *for whom* questions, as well.

This section discusses only the first of three main variants of the "for whom" question: people who take up the offer of treatment, subgroups defined by characteristics measured prior to assignment, and subgroups defined by characteristics that may be themselves affected by treatment or the offer of treatment. The second for whom question is addressed at length in Chapter 7 in this volume, and the third is fundamentally connected to the questions of mediation and moderation we discussed above.

With the exception of BOND Stage 1, in SSA's demonstrations, assignment to a treatment or a control group means treatment group members are offered access to the treatment, and control group members are offered access to whatever services are business as usual. Some members of the treatment group could choose not to use the treatment, and some members of the control group might obtain the treatment (or something that closely approximates the treatment), both forms of "crossover" between arms of the demonstration. In the presence of any crossover, an evaluation using assignment measures the impact of the *offer* of the intervention, termed the ITT impact. ITT evaluation designs give overall average effects that we would expect to see when participation in a program is voluntary.

That is, the ITT evaluation estimates average effects of the offer of an intervention (such as a benefit offset or extra counseling), which is exactly the target quantity of policy interest when we are thinking of a new policy that offers that intervention broadly. But in many situations, we imagine making the intervention mandatory or universal, and then we would like to know the effect of the intervention itself, not the offer of it.

Going beyond the ITT estimates that have been the focus of SSA demonstrations has the potential to provide additional useful information on the effectiveness of the treatment. To assess the effect of the treatment on those who would not have received it under the status quo, evaluators use exactly the same data to estimate what has been termed the TOT effect.[7]

The TOT estimates, and the effect they are estimating, are of greater policy interest when the share of the sample taking up treatment (or the composition of that group) can itself be manipulated via other interventions, or when variation in treatment effects is thought to be tied to participation. In the first case, if there is a large TOT effect and a small average treatment effect, then we can increase the share of the sample getting the benefit of treatment via the other interventions and improve overall impacts. In the second case, it might be true that average treatment effects are small but that there is a population for whom they are large, and those individuals seek out the treatment.

Ignoring the difference between TOT impacts and average impacts can produce incorrect conclusions when examining the evaluation in a demonstration. For example,

---

[7] Other terms used to describe this effect are the local average treatment effect (LATE), the complier average causal effect (CACE), or the average treatment effect on the treated (ATET).

a service used by a vanishingly small share of participants could be judged ineffective simply because its effect is averaged together with zeros for those not participating, yet it could be very effective for that small share of the sample who take it up. The gap between the TOT and average impact could illuminate that.

For example, as part of the PROMISE demonstration, federal partners identified a set of core services that could achieve the desired results on educational attainment and employment, and thus required the PROMISE programs to include those services. Not all SSI recipients who were assigned to the treatment accessed the core services, and some members of the control group received the core services. Consequently, the PROMISE ITT analysis will likely understate the effectiveness of the core services that federal partners deemed important. SSA's Project NetWork and AB demonstrations are other examples where either some members of the treatment group did not use the core services, some members of the control group obtained the core services, or both. As we will describe below, in those instances, the ITT estimates from the impact evaluation likely understate the effectiveness of the core treatment components.

The Oregon Health Insurance Experiment evaluation includes a good application of estimating TOT (Finkelstein et al. 2012). A group of uninsured low-income adults in Oregon were randomly assigned to be eligible to apply for Medicaid coverage, and a year later they were about 25 percentage points more likely to have Medicaid, compared to those not randomly selected to become eligible. The study estimated the ITT effects and found that those selected to the Medicaid-eligible group exhibited statistically significantly higher health care utilization (including primary and preventive care, as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the group not selected.

The study also estimated the causal impact of insurance among the subset of individuals who obtained insurance due to being randomly assigned to the Medicaid-eligible group and who would not have obtained insurance without being randomly assigned. These TOT estimates were approximately four times larger than the ITT estimates. These estimates provide causal effects of the Oregon Medicaid program itself on outcomes, as opposed to the effects of the offer to obtain health insurance through the Oregon Medicaid program. This information is particularly useful to policymakers, as it provides an estimate of the impact of the Oregon Medicaid program on a variety of outcomes. It could be used as a basis for conducting a cost-benefit analysis of the Oregon Medicaid program for those participating.

The evaluation of the Oregon experiment also provides a good example of the type of assessment that evaluators should conduct when developing TOT estimates. First, the study considers various measures of health insurance during the study period and describes the corresponding implications for the TOT estimate. Second, it provides an assessment on the additional identifying assumption that there is no effect, on average, on the outcomes studied of being randomly selected to access the Oregon

Medicaid program that does not operate via the experiment's impact on insurance coverage.

The authors identify two potential violations. First, the event of being selected to access the Oregon Medicaid program might have direct effects on the outcomes studied. Finkelstein et al. (2012) convincingly describe the reasons that this is unlikely to be the case for outcomes one year after selection into the study. Second, individuals who apply for public health insurance might also be encouraged to apply for other public programs for which they are eligible, such as food stamps or cash welfare. Therefore, the mechanism behind the effects on outcomes might be partly attributable to participation in these other programs and not entirely to the Oregon Medicaid program. The authors have access to data to assess the extent to which this might be the case, and their assessment indicates that it is very unlikely. The evaluators' careful analysis of potential threats to the validity of the TOT estimate promotes confidence in the results of their evaluation. Their pre-specified analysis plan (Finkelstein et al. 2010) uses the control group data to guide analysis except in one key respect: they examine the first stage to see the effect of offer on insurance status to see whether TOT estimates will prove useful.

The AB demonstration is the only SSA demonstration project that develops TOT estimates to provide policymakers with information on the impact of the TOT (Weathers and Bailey 2014). Similar to the Oregon experiment, the AB project examined the impact of a health insurance package on a variety of health outcomes, and it targeted beneficiaries who did not have health insurance coverage at the time of random assignment. Approximately 35 percent of those in the control group reported that they obtained health insurance within one year of random assignment. Therefore, control group members were able to access health insurance similar to the health insurance received by the AB group and the AB Plus group. The authors of the study show that the ITT estimates indicate that assignment to either the AB or AB Plus treatment groups resulted in statistically significant improvements to self-reported health at one year after enrollment into the health insurance program, positive effects on physical and mental health measures one year after enrollment, and no statistically significant effects on mortality during the study period. Estimated TOT effects on the health and mortality measures are approximately 1.5 times larger than the estimated ITT effects. These larger estimates point to the impacts we would expect from universal coverage, rather than the offer to sign up. The effects of the offer of coverage are of direct interest, both in the Oregon experiment and AB demonstration, because Oregon wants to know how many would sign up when offered coverage and what improvement in average outcomes might be, and SSA has similar interests. But the effects of the coverage itself was of pressing national interest at the time of the Oregon experiment, since a national policy conversation was underway about expanding coverage, possibly reaching universal coverage. That is, Congress and the public wanted to know the effect of insurance, not the effect of an offer of insurance.

Presumably, the AB demonstration could also lead to a policy change that affected all applicant's insurance status.

However, a limitation of the authors' TOT analysis is that it did not include the careful assessment of potential violations to the required assumptions, as was done in the Oregon experiment. The limitation occurred because the TOT analysis was not part of the original evaluation design, and the information necessary to conduct the supporting analysis of the necessary assumptions was not part of the data collection plan—which highlights the importance of including plans to estimate TOT in the design phase of a demonstration.

Project NetWork is another example of where the TOT estimate could have been useful to policymakers. Those individuals assigned to the treatment group met individually with a case or referral manager who arranged for rehabilitation and employment services, helped clients develop an individual employment plan, and provided direct employment counseling services. Volunteers assigned to the control group could not receive services from Project NetWork but remained eligible for any employment assistance already available in their communities. The evaluation documented the rehabilitation and employment services that each of two groups received, as shown in Exhibit 3.1. Notably, the differences in the receipt of the employment services was relatively small, and the estimated impact on earnings was small, as well.

Exhibit 3.1. Receipt of Education, Training, and Rehabilitation Services from the Project NetWork Follow-Up Survey Sample

Service Received since Random Assignment	Control Group (%)	Treatment Group (%)
Job search assistance	14	21***
Business skills training	6	11***
Job-related training	10	12
Other rehabilitation/training	2	1
Life-skills training	6	6
Occupational therapy	4	4
College classes	10	8
Assessment of work potential	17	27***
Physical therapy	23	23
Psychological counseling	38	41
Any service	69	75**

Source: Kornfeld and Rupp (2000, Table 6).

**Difference between levels for treatment and control groups is statistically significant at 5 percent level.

***Difference between levels for treatment and control groups is statistically significant at the 10 percent level.

The primary effect of Project NetWork was that it increased earnings in the first two years after random assignment by about $220 dollars each year. This was considered a small impact in the report; and it is not too surprising given that 69 percent of the control group members were able to access services that would facilitate

a return to work. Using the information on the difference in rates of those who obtained "any service" shown in Exhibit 3.1, across those assigned to treatment and control, a rough approximation of the TOT would indicate a much larger effect when compared to the ITT estimate, about $3,666 per year (or $220/0.06) in the first two years after random assignment. This estimate is based on a number of assumptions, and a much more careful analysis would be needed to support it, and in particular to construct a confidence interval around the estimate. However, this is the type of information that is important to policymakers in that it shows that the services had a much more substantial effect on individuals who would not have received the services without Project NetWork.

The ITT is always of interest in a demonstration, as is the rate at which groups receive the intervention (or receive services or are exposed to policy environments that are similar to the intervention). But specifying potential TOTs in the demonstration design report and developing a plan for estimating TOTs can increase the information that a demonstration produces. The information could be informative for assessing effects under an alternative service environment where access to the intervention's services or similar services is much more difficult. It could also provide information on how to better target the services to reduce the costs of the program without reducing the benefits of the program. Finally, it could provide policymakers with the evidence needed to make more-informed decisions on a program or policy. For example, policymakers might view the value of a program differently if they understood that it substantially increased outcomes for a relatively small group instead of producing a minor impact on a relatively larger group. Consequently, TOTs have a substantial potential for establishing evidence that could influence decisions on implementing a program or policy.

Group-specific TOTs and group-specific first-stage effects of the offer of treatment on take up of services would typically be of greatest use in devising whether targeted service offers would produce larger impacts than would a general offer. The first-stage effects inform us about likely costs related to increased service use in each group, and the TOT estimates tell us about the impacts on those who get the services as a result of the offer. In some cases, a policymaker might be interested only in the overall average impact of a policy rolled out to all individuals at once, and might believe take-up cannot be manipulated, in which case the TOT is irrelevant: a simple comparison of means captures both differential services and average impacts. But we assert the policymaker should be interested in TOT as well. In particular, if that same policymaker could learn that an overall average treatment effect close to zero were a blend of large positive impacts on one subgroup and large negative impacts on another, and that costs of the offer differed substantially across groups, then the policymaker should clearly value that information.

At minimum, an evaluation associated with a demonstration should report estimates of the average treatment effect both unconditionally and regression-adjusted, where applicable, and both the ITT estimate and the most relevant TOT estimate.

Estimates should be paired in all cases with their standard errors including multiple statistically significant digits (not simply stars or a *p*-value, and no standard error should ever be reported as 0.00). Finally, the evaluators should report the estimated difference between the ITT and the TOT estimates, and the standard error on that difference (or at least the result of a test that they differ). Yet to the best of our knowledge, no demonstration described in this volume has met this standard.

Designing a Cost-Benefit Analysis

The appropriate treatment effect (or effects) to estimate in the evaluation tied to a demonstration depends on a specific policy question (or set of questions), but the treatment effect is only half the story: the costs and benefits that accrue from a treatment effect will determine optimal policy responses. The typical cost-benefit analysis counts up costs and benefits associated with impacts and computes a net benefit in moving from the control to the treatment condition, from a particular perspective. That is, we can imagine computing net benefits for treatment group members, for the government balance sheet, or for society at large. This last perspective is especially important when we consider a social insurance program, as those individuals who take up benefits are not the only ones who benefit from the existence of (or design of) the program. A final point we will return to is that the demonstration itself has costs and benefits, which might be estimated as part of the demonstration.

Appropriate costs also include opportunity costs, such as the value of time spent at work or the opportunity cost of government funds expended (which for an evaluation of a policy or program would exclude the costs related to the demonstration itself). This simple point is not universally acknowledged in past SSA demonstrations; for example, Decker and Thornton (1995) do not compute the opportunity cost of government funds and assign a negative value to the non-work time of individual participants. The measurement of the full economic value of any difference in outcomes involves thorny problems in both the empirical and theoretical approaches. No single solution is likely to satisfy all readers. Another feature that is often overlooked (e.g., in the Transitional Employment Training Demonstration, AB demonstration, and others) in computing costs and benefits is the sampling variability in a demonstration. Any measure of costs or net benefits should include a confidence interval, which raises a different set of empirical and theoretical issues. We like to think of a confidence interval as arising from repeated sampling from a population, ideally with independent draws, but the right inference for the net benefit calculation when we want to understand the deep structural issues in policy is some super-population of possible populations. That is, we should report a confidence interval as if we could somehow repeatedly run a demonstration on the full population. At minimum, we want to account for correlations of estimated differences in various costs and benefits when we construct the confidence interval for their sum.

Estimating even individual appropriate costs and benefits, by "payer" or societal component, is easier said than done. Comparing the sum of benefit payments and administrative costs across treatment and control groups during the period of the evaluation seems an easy first approximation to the government's perspective on net benefits, but it doesn't take account of any spillovers onto other programs, difference between short-term and steady-state responses, and discounting of future net benefits (i.e., an average over several years is not the long-term present net value). Spillovers onto other government programs or tax collections are one form of externality, but there could be many: there could be effects on participants' extended families, or future generations. Even just accounting for costs to one government program, we need to discount costs in future years to the present, which involves difficult and even controversial choices that can have real consequences for policy design.

The state of the art in such cases is to compute costs using several different approaches and report each, as in Gubits et al. (2018a/b). Because reasonable people can disagree on the premises for each approach, it is safer to report results for many possible approaches, from different perspectives. When estimates are reported for each option, for example no discounting or using two different discount rates, readers get a sense of the sensitivity of the result to alternative choices. Likewise, authors can account for the value of time out of the labor force using several different methods, reporting each set of estimates.

This is also the typical solution to the thorniest problem in computing costs and benefits from a societal perspective, which concerns the distribution of benefits and costs. Because costs might be higher for some people and benefits higher for others, any given policy can create "winners" and "losers." Balancing across these groups requires social welfare weights. Note that a strictly utilitarian approach of equal weight for each individual is itself a choice of weights, and reporting for several options is one way forward. For example, Gubits et al. (2018a/b) report different possible net benefit estimates for society, using equal weights or alternatively higher weights on beneficiaries, who have very low incomes and therefore a presumed higher marginal utility of money.

Hendren (2020) provides an alternative framework for comparing costs and benefits of alternative interventions, dealing both with the opportunity cost of public funds and with the distributional weights applied to winners and losers. The Hendren approach uses the ex-ante marginal value of public funds (EMVPF), which relies on comparing to the social welfare weights that are implied by the tax schedule. This works well if income is an index that identifies the most important source of variation in social marginal utility, because we can imagine transferring value to individuals at different points in the income distribution with a range of treatments that includes novel tax policy as one option. Hendren and Sprung-Keyser (2019) estimate a modest EMVPF for SSDI (similar to effecting transfers via the tax code) by relying on estimates of effects due to marginal changes in SSDI benefit generosity (Gelber, Moore, and Strand 2017), SSDI administrative law judge leniency (French and Song

2014), SSDI medical examiner leniency (Maestas, Mullen, and Strand, 2013), and changes in veterans' disability compensation (Autor et al. 2016).

The Hendren framework is a useful way of avoiding explicit treatment of opportunity costs, and compares to the tax and transfer system to pin down the *relative* social marginal utility of transfers to individuals at different points in the income distribution. In the absence of that framework, one has to account for opportunity costs using an alternative approach, such as valuing the social costs of increased government costs using a markup related to the deadweight loss or excess burden of taxation (as is done in Gubits et al. 2018a/b). But this novel framework is unlikely to be a magic bullet for SSA demonstrations and policy for two reasons: income is not the only thing that matters in disability policy; and SSA takes a long view, whereas tax policy can change substantially over time and therefore implied social welfare weights can change, as well. That is, the Hendren framework also relies on many assumptions, that could prove too far a reach for SSA demonstrations to justify using as a single organizing principle in cost-benefit calculations.

Building a variety of approaches to cost-benefit analysis into future demonstrations seems key to future-proofing the results. One important aspect of SSA demonstrations is that they should also provide the information on the government balance sheets needed to evaluate interventions. Government balance sheet effects are needed (e.g., by the Congressional Budget Office or the SSA Actuary) to evaluate costs of any proposed policy changes tied to specific legislatively authorized forms of spending. That is, even if a policy change produces a large social gain, if it draws down an agency's budget, then Congress would have to appropriate funds to make up that shortfall.

IDENTIFYING AND ACQUIRING THE DATA NEEDED FOR EVALUATION

Evaluation plans are meaningless without the data necessary to carry out the evaluation. A demonstration design report must include a description of the data needed for evaluation purposes; and the logic model provides the overall architecture for the use of those data to measure inputs, outputs/activities, and outcomes. This section describes three sources of data used for demonstrations, and opportunities to improve the information drawn from each of these sources.

Surveys have the advantage of collecting customized information that is not available from administrative data sources, but can suffer from low response rates, recall bias, or other biases inherent in survey data. SSA's administrative data has the advantage of being available for all demonstration participants who can be matched to the data. The administrative data is not subject to recall or social desirability bias (though it can have other kinds of errors, for example when multiple people's earnings are attributed to one Social Security number), and can provide more objective data than survey data. Administrative data, however, has the disadvantage of being limited in scope and it can become available only after a lag (e.g., tax returns might be

available years after income is supposed to be measured, and then might not measure all forms of income). Administrative data from other federal agencies have similar advantages to SSA's administrative data and can fill in some of the holes in SSA's administrative data, but also has the disadvantage of being limited in scope to the programs for which those data are collected.

Therefore, it is important for the data collection plan to use a combination of these data in a manner that takes advantage of each source's strengths and that can provide information on the potential for bias in survey data.

Use of Surveys

Surveys of demonstration participants and other respondents are often the most expensive part of a demonstration, so maximizing the value of this type of data source is a key concern for SSA. Data on a wide variety of potential moderators or mediators of treatment effects cannot be obtained from administrative data. The main use of survey data in SSA demonstrations has therefore been to collect information that is not available in administrative data, such as knowledge of program rules among demonstration participants or demographic information useful for subgroup definitions. For example, race and ethnicity are not reliably measured in administrative data (Martin 2016). Exposure to environments that cause future disability is also not measured by administrative data. These types of exposure could be useful in understanding how impacts of interventions are related to individual characteristics (improving targeting, if impacts are heterogeneous). They also could suggest wholly new kinds of interventions that reduce disability prevalence and program entry rates, rather than promoting work and program exit.

It can be very helpful to build mini-experiments into survey collection as part of a demonstration to learn more about how best to conduct surveys. For example, demonstrations' survey efforts can help inform whether it is helpful enough to send a letter before contacting a potential respondent to justify the extra cost and time (Vogl et al. 2019). Interviewer effects can bias answers or produce unacceptable variation in response rates (Lavrakas, Kelly, and McClain 2019), and the race and ethnicity of both respondents and interviewers seem especially important to examine (Holbrook, Johnson, and Krysan 2019). For example, suppose a hypothetical evaluation found that an intervention tested in a demonstration had impacts only on the understanding of program rules for White non-Hispanic respondents, but all the survey interviewers were White non-Hispanic. In this situation, we might not trust the finding, as we would like to know that findings are robust to the race and ethnicity of the staff conducting the evaluation.

Traditionally, surveys could reach large swaths of the population (e.g., to conduct polls about an upcoming political election, researchers could send mail to address lists or randomly dial phone numbers). But response rates in random-digit dial or mail surveys of the general population have fallen into the single digits over the past decades, with survey firms struggling to achieve double-digit response rates (Kennedy

and Hartig 2019). SSA demonstrations regularly have 80 percent response rates, as in the recent PROMISE surveys (Mamun et al. 2019), but it is important to note that often their population is pre-selected by virtue of volunteering to participate in the demonstration. Soliciting the entire pool of SSDI beneficiaries or SSI recipients produces much smaller responses, typically in the single digits. For example, an 84 percent response rate for a survey of BOND Stage 2 participants at the 12-month mark is less surprising when we consider that the frame includes only the 5 percent of eligible SSDI beneficiaries who already volunteered at an earlier point. In the National Beneficiary Survey (NBS), response rates have fallen in recent rounds relative to earlier rounds:

> 14 percent of households contacted refused in Round 5 compared to 12 percent in Round 4…approximately 13 percent of the sample members were not located at the end of data collection in Round 5, compared to 9 percent in Round 4 [and] contact information was invalid for [five of every eight] beneficiaries in the sample[; placed] more calls on average to complete an interview than…in the prior Round 4 NBS (36 percent versus 31 percent) [and saw more] "noncontact" status (that is, repeated attempts that end with an answering machine or no answer at all)—13 percent of the sample compared to 9 percent in Round 4. (Skidmore et al. 2017, 5)

There are four main advantages to SSA demonstrations using surveys, and the first is that we can improve on the usefulness of the survey instrument for collecting data in the demonstration at hand. This often uses "adaptive design" where features of the survey can be modified on the fly, which is especially easy in web-based surveys (Kunz and Fuchs 2019). The second is that surveys themselves can embed informational "nudge" experiments that can be analyzed for years after the demonstration is over; that is, modules can be randomly varied across respondents to deliver an intervention in the form of information. The third related advantage is that a well-designed survey experiment can provide information relevant to treatment effect heterogeneity (see Chapter 7), moderation, and mediation (Tipton et al. 2019). The fourth advantage is that a randomized incentive can allow the demonstration to extrapolate evaluation findings to a larger population much more easily and plausibly.[8] Unfortunately, Office of Management and Budget approval is by no means guaranteed for this last feature, or for modifications to an ongoing data collection effort, and more work needs to be done to motivate these innovations. Furthermore, adding a nudge in

[8] For example, in POD, the six percent of participants who exit the survey create a real bias as at least a third of them are better off bypassing POD rules, meaning they are responding to the financial incentives to work by exiting the demonstration. If randomized incentives were included in the design, the precise nature of the bias due to exit could be estimated using that random variation in incentives to remain.

a survey also represents a modification to the intervention, which could change the interpretation of findings in the overall evaluation.

Because surveys are expensive, often accounting for the bulk of evaluation-related costs of a demonstration, there is often a desire to collect data only via already available administrative data. However, each data source has a separate value, and the two together give us an extra insight into variables measured in both.

Use of SSA Administrative Data

SSA administrative records—collected as part of the normal administration of Social Security programs—are an important source of data for demonstrations. The data are available for all treatment and control group members, and they are not subject to the limitations inherent with survey data such as survey non-response, item non-response, recall error, or other forms of systematic measurement error.

In the past, access to these data required an SSA employee to request the data through a "finder" process that would be conducted by another authorized SSA employee. The finder process uses a file that contains unique personal identifiers to extract data from an SSA administrative data file. SSA's demonstrations usually require data from several SSA administrative data files. The primary sources are the Master Beneficiary Record for information on SSDI beneficiaries, the Supplemental Security Record for information on SSI recipients, the Master Earnings File (MEF) for information on earnings and other income (Olsen and Hudson 2009), the 831 file on applications and determinations, the Waterfall file on continuing disability reviews, and the Numident on birth, emigration, and death dates (and legal status at entry into the United States). A separate finder process is required to obtain data on a demonstration's participants from each of SSA's administrative data files. The evaluation team then needs to merge these files and convert the administrative files to data files that are suitable for research and evaluation purposes. This process can be very cumbersome, as illustrated in the documentation describing the construction of the files originally developed for the evaluation of the Ticket to Work program (SSA/ORDP/ORDES 2020).

Though rare, there have been instances where the finder process produced a file that was either incomplete or contained errors. In such instances another request must be submitted and executed. Thus, although the finder process is useful, it has limitations and there is room for speed and accuracy improvements.

To overcome many of the limitations involved with the finder process, and to make the data more accessible for research and evaluation purposes, SSA invested in the development of the Disability Analysis File (DAF). The DAF is a set of files containing SSA administrative data on federal disability beneficiaries, culled from a variety of SSA sources commonly used in program evaluation and research. The DAF includes longitudinal data on program participation, employment activity, and benefits for adults who have received SSDI payments and children and adults who have received SSI benefits in any month since 1996. The DAF also includes detailed

documentation on the data, which is organized in a way that is relatively easier to use than data produced from the finder process. Consequently, the DAF is an important resource to the research and evaluation community.

Though the DAF is an advancement over the finder process, a drawback of the DAF is that it is developed once per year. As of November 2020, the DAF18 files are available to researchers, covering data through calendar year 2018. Another drawback of the DAF is that though it contains a vast amount of data, there can be instances where the data needed for an evaluation might not exist in the DAF. For example, SSA administrative data on whether the individual is identified as homeless are not available in the DAF, so the DAF was not a sufficient data source for the evaluation of the Homeless with Schizophrenia Presumptive Disability Pilot (Bailey, Goetz Engler, and Hemmeter 2016).

One potential opportunity to improve the timeliness of obtaining data for research and evaluation purposes is to leverage the data and computing capabilities within SSA's Enterprise Data Warehouse (EDW) to update the DAF more regularly (e.g., monthly). The EDW contains data not currently in the DAF, such as the information needed to identify disability applicants who are documented as homeless at the time of application for disability benefits. For SSA's demonstrations, the integration of the DAF into the EDW has the potential to provide policymakers with more timely information on key findings from the demonstration evaluations.

Uses of Administrative Data from Other Government Sources

Another opportunity to improve the information that demonstrations produce is to make greater use of administrative data from other government sources, including other federal agencies and state sources. Indeed, this is a goal of the Foundations for Evidence-Based Policymaking Act of 2018 (Hahn 2019). Additional federal sources include health services data from the Centers for Medicare and Medicaid Services (CMS) and data on earnings from the Office of Child Support Enforcement or the Internal Revenue Service (IRS).

SSA's demonstrations have used a variety of these data in past demonstrations. The State Partnership Initiative (SPI) demonstration used state Unemployment Insurance (UI) data and SSI administrative data for one site (New York). The Benefits Entitlement Services Team (BEST) used data from the Veterans Benefits Administration on Veteran's Disability Compensation claims. All the demonstrations reviewed aside from the MHTS used both administrative and survey sources of earnings data (MHTS did not collect administrative data on earnings and employment, but collected monthly data from surveys). However, the demonstrations have not used these rich data as effectively as one might expect. In particular, each data source has very different sources of error, and having both survey and administrative data available offers the evaluation the opportunity to improve on estimates using either one or the other. Instead, each SSA demonstration reviewed reports separate estimates

for separate data sources instead of matching like concepts in disparate data to better measure the underlying concepts, for example employment, income, or health.

Centers for Medicare and Medicaid Services. CMS data on Medicare and Medicaid is important for demonstrations where a potential outcome is a reduction in long-term medical care expenditures. For example, the AB demonstration logic model identified reducing reliance on Medicare and Medicaid as an ultimate outcome of the AB health insurance package (Weathers et al. 2010). Similarly, the MHTS and SED could reduce reliance on Medicare and Medicaid (Frey et al. 2011).

Unfortunately, challenges with establishing an agreement between SSA and CMS have prevented use of that data for research and evaluation purposes. The provisions within the Evidence Act could provide both agencies with incentives to engage in a data sharing agreement that would strengthen the evidence on how the earlier provision of health services might reduce reliance on these programs. However, the Evidence Act did not provide any new authorities or funding, and it did not address issues related to routine uses and the various privacy laws that hinder the use of data. Data that are routinely shared across government agencies for operational purposes, such as checking ongoing eligibility for programs, is sometimes specifically prohibited from use in research, even though the potential harms are often much lower in research use than in operational use.

Office of Child Support Enforcement. Another useful administrative data source is data from National Directory of New Hires, which is a national database of wage and employment information that consists of three files: new hires, quarterly wages, and UI. The new hires file contains information on all newly hired employees, including the date of hire. The quarterly wage file contains quarterly wage information on individual employees that is submitted by state workforce agencies or federal agency records. It contains a separate record for each job. The UI file contains UI information on individuals who received or applied for unemployment benefits, as reported by state workforce agencies. The states only submit claimant information that is already contained in the records of the state agency administering the UI program. These three files provide more detailed information on employment than what is available in SSA's MEF, which has annual data on earnings. Therefore, their information would provide a more detailed picture of work behavior during the course of the year than is available from SSA.

Internal Revenue Service. A third source of data that would be useful is data from federal tax returns, including 1040 forms and information returns, which have been used by others to develop family income measures (Chetty et al. 2017). Tax returns contain a wealth of information, including on college attendance (via 1098-T forms), non-wage sources of income such as interest and dividends, and moonlighting or other kinds of self-employment income (via 1099 forms). A panel of tax returns can provide information over long stretches of time on family formation or dissolution and migration across states, without the need to follow respondents to new addresses and survey them. The data have the potential to be particularly useful for the PROMISE

demonstration, where the conceptual framework specifies higher family income and economic well-being as key long-term outcomes (Fraker, Carter, et al. 2014).

Given the potential limitations of using survey data to measure employment and earnings, it seems likely that surveys might not be an ideal source for measuring family income. However, there are some forms of income not captured in administrative data, such as moonlighting or unreported self-employment income, so pairing both a survey and administrative data source is frequently the most desirable option (if also the most expensive option). This is the approach taken by most of the SSA demonstrations reviewed in this volume, including BOND, which reports "no meaningful effects on survey-measured outcomes" constructed from Stage 1 and Stage 2 survey data in the final report, and refers readers to supplementary reports for estimated impacts; "results are presented in Hoffman et al. (2017), Gubits et al. (2017), and Geyer et al. (2018)" per Gubits et al. (2018a, 63).

None of the demonstrations we reviewed exploit the multiple sources of data available to explore the nature of measurement error in the different sources of data. For example, a survey measure of health might use a short form designed to measure underlying health outcomes and report that the short form has been validated, and then report impacts on mortality, but never compare the two measures of health. BOND used the 36-month follow-up survey in Stage 1 for several measures of health and reported that "estimated impacts on these measures vary in sign and are generally of negligible size and statistically insignificant" (Gubits et al. 2018a) and then reported mortality differentials in the report's Appendix F (see Gubits et al. 2018b).

Using a principled framework such as that proposed by Kapteyn and Ypma (2007) can substantially improve over reporting separate estimates for outcomes based on survey and administrative data sources. Understanding the discrete properties of different data source is also crucial to using these various sources. For example, finding large impacts in survey measures of earnings but not administrative reports suggest third-party reported income responds less than unreported income, but larger impacts in administrative data could suggest some reclassification of income. Objective and subjective measures of health may respond quite differently, and the interpretation of any difference is not straightforward. But when net earnings and employment are key outcome measures, as they have been in every SSA demonstration we reviewed, using a principled measurement error framework can only add value to the evaluation of a demonstration's findings.

EXPANDING THE DISSEMINATION OF FINDINGS TO STAKEHOLDERS

A well-executed dissemination strategy is an important ingredient for a successful demonstration, as a successful demonstration is one that generates information that is used to inform policy. A missed opportunity exists when results from a demonstration project are described in a report that does not reach key stakeholders. When this occurs, the return on the substantial investment in the demonstration is suboptimal. Although publishing findings in academic journals can increase the credibility and

visibility of project findings, it generally is not a sufficient means of reaching key stakeholders. In this section we describe strategies to communicate demonstration findings to policymakers and stakeholders to position those findings for use.

Engage Stakeholders Early and Often

A useful starting point is to engage stakeholders beginning with the development of a demonstration and continuing throughout implementation of the project. Stakeholders can have unique insights into the information needed from a demonstration, and incorporating their input into an evaluation plan will help ensure that the demonstration addresses their needs.

The initial development of BOND provides a good example of engaging a broad group of stakeholders in the development of the demonstration. The demonstration was required as part of the Ticket Act, and SSA sought the advice of the Ticket to Work and Work Incentives Advisory Panel in the early stages of development. The panel represented a cross section of experience and expert knowledge as recipients, providers, veterans, employers, and employees in the fields of employment services, Vocational Rehabilitation, and other disability-related support services. The panel developed an advice report on the demonstration based on relevant documents and testimony, including several SSA reports considering SSA's draft evaluation plan for the $1 for $2 offset program. The panel also obtained information by conducting an Experts Roundtable (Washington, DC, November 16, 2001) and from public comments made before and after the roundtable. The panel's final advice report (*US Ticket to Work and Work Incentives Advisory Panel 2002*) included eight recommendations, and SSA incorporated almost all of them in the BOND's design. The report was also sent to members of Congress.

This strategy for developing BOND was important for several reasons. First and foremost, it strengthened the design of the demonstration and its evaluation plan. For example, the panel recommended deferring the evaluation of induced entry into the program, and recommended focusing the demonstration on current beneficiaries. The panel also recommended the use of the following employment supports to be in effect for both the treatment and control groups throughout the duration of the demonstration:

- Access to local community-based benefits planning services (or their reasonable equivalent);
- Access to local community-based protection and advocacy services (or their reasonable equivalent);
- Access to responsive local work incentives specialists (or their reasonable equivalent) within SSA; and
- Access to ongoing, understandable information on the treatment and its interaction with other programs and services administered by SSA.

Second, the panel's strategy for obtaining information from a broad array of stakeholders proved to be effective in raising awareness about the demonstration.

Third, by considering and incorporating the panel's recommendations into BOND, SSA demonstrated a commitment to incorporating a diversity of views into the development of the demonstration. Finally, the strategy helped develop the momentum behind the demonstration to move it into the implementation phase. Unfortunately, the planning process for BOND stretched out to nearly a decade, and while that time was well spent, policymakers outside of SSA became impatient for results, and have restricted SSA's planning time in some recent demonstrations.

SSA has led other positive developments in planning, which could serve as models for other government agencies planning demonstrations. For several recently proposed demonstrations, SSA has convened technical expert panels (TEPs) that provided SSA with information necessary to define the scope of a demonstration and to develop an evaluation strategy. As one example, SSA used a TEP to help define demonstrations related to post-entitlement earnings simplification and shape the future Exits from Disability Demonstration (Gubits et al. 2019). The various TEPs included members from academic research institutions; federal government agencies outside of SSA; private non-profit policy advocacy and analysis organizations or non-profit service providers; private businesses; and independent consultants. The panels have assisted SSA with developing research questions, intervention specifications, implementation strategies, and evaluation designs to ensure that demonstrations generate the evidence SSA needs to inform policy decisions.

The TEPs provide SSA with objective review of potential demonstrations and independent recommendations regarding what SSA might study. Though the Post-Entitlement Earnings Simplification Demonstration TEP's activities were more limited in scope compared to the work of the BOND panel, the TEP report provides SSA with a strong foundation for the development of the demonstration. SSA also convened a TEP for the PROMISE demonstration and the Work Incentives Simplification Pilot, and it internally put together a TEP for both SED and POD. These TEP findings represent a middle ground between implementing a demonstration without external guidance and feedback, and the decade-long planning period involved in BOND. It seems unlikely that any new demonstration would be allowed to explore options for a decade, given the criticism of BOND's slow startup. The foreshortened preparations for POD and the Retaining Employment and Talent After Injury/Illness Network demonstration necessitated by the authorizing legislation might reflect the pendulum's swing toward hasty implementation. A deliberative TEP offers a useful compromise, and demonstrations mandated by Congress should allow for a deliberative planning process to improve the demonstrations' usefulness.

Importance of Disseminating Interim Results Early

The final evaluation for a demonstration often occurs several years after its initial implementation because participants need time to respond to the intervention, evaluators need time to collect the data necessary for a final evaluation, and then evaluators need time to process the data and draft a final report. Stakeholders have

expressed frustration at how long it takes to initiate demonstrations and obtain findings from them.

One opportunity to improve the value of a demonstration project to stakeholders is to disseminate key findings before the completion of a final report. SSA has done this to some extent with interim reports produced for the PROMISE demonstration and BOND, but there are opportunities to disseminate key findings in a more accessible and timely way. This could be done by developing and disseminating a series of two- to three-page briefs throughout the course of a demonstration to highlight key findings to date, in particular for inputs and outputs, before impact estimates are available. These briefs could be disseminated by SSA on its website, sent via email to stakeholders, and highlighted on SSA's social media outlets. Such briefs would not be a substitute for a thorough evaluation report, but would provide stakeholders with more timely information on results of interest and keep them engaged in the demonstration's activities.

The Office of Evaluation Sciences (OES) within the General Services Administration provides a good illustration of this approach. In addition to developing detailed evaluation reports, OES produces two-page "abstracts" that describe the evaluation and its key findings. A good example is an abstract of work OES conducted with SSA to encourage SSI recipients to self-report wage changes (GSA/OES 2019c). OES disseminates these abstracts on its website and highlights them in its social media blog. The abstracts provide stakeholders with the clear and concise information they need to make decisions. Abstracts can be completed prior to the release of the more detailed evaluation reports. The OES model could work for SSA, with the FLM approach to report on participation inputs and early contrasts in outputs. In some cases, proximal outcomes and impact estimates also could be promulgated to build interest in the demonstration's eventual findings.

Very few people read detailed technical reports, and fewer still read academic papers. For example, the World Bank spends a large fraction of its budget on knowledge diffusion, but "more than 31 percent of [World Bank] policy reports are never downloaded," and nearly 9 in 10 policy reports were never cited (Doemeland and Trevino 2014). More broadly, nearly 6 in 10 academic articles are never cited more than once, and 44 percent are never cited (van Noorden, Maher, and Nuzzo 2014). Of course, citation is only one measure of influence, and many might read a report but never cite it. Nevertheless, these statistics indicate that reports designed to be downloaded are often never downloaded.

It could be that alternative mechanisms for distributing findings would be more effective, but research on this is scant. Regarding public health research "social media dissemination is significantly positively associated with more downloads and eventual citations" but "it is unclear whether tweeting science influences, or is merely correlated with, citations" (Brownson et al. 2018). It could prove useful in a future multisite evaluation to randomize strategies for disseminating findings from each site in order to learn more about which actually get the word out. SSA's Office of

Communications could test alternative communication strategies to learn which achieves the greatest reach. Doing so would add value to future demonstrations.

BROADENING THE USE OF DATA FROM THE DEMONSTRATIONS TO INFORM PROGRAM AND POLICY DEVELOPMENT

The information that demonstrations produce should not languish in a final evaluation report. There are several opportunities to make use of data from a project to build a stronger evidence base for policymakers to use when deciding whether to implement a new policy or program. In this section we describe three: (1) using qualitative findings to improve on theoretical models; (2) making data available for additional analyses; and (3) conducting meta-analyses to learn more from past demonstrations.

The findings of interest in a demonstration are not only distal impacts or changes in outcomes, though often readers take away only one top-line finding on a final outcome. As discussed in Chapter 9, enrollment rates, for example, can indicate the level of interest and the response to an intervention should it become national policy. Similarly, contrasts in service use between treatment arms can indicate the likely reach of a future national policy relative to current law. That is, say an intervention includes a service used by half of the treatment group members, who then stop using a very similar service used by most of the control group. That finding implies the intervention might produce a change in the type of service used but no change in amount of services; a national policy implementing that same intervention might simply be reassigning the responsibility for service delivery.

These findings are all useful in measuring the steady-state effects of an actual policy change. But as we highlighted above, using a logic model, the analyst can learn more than just how one policy versus another compares. In particular, the theoretical model used to build the logic model could prove incomplete in light of the findings, if high-quality inputs fail to lead to outputs, or high-fidelity outputs fail to yield hypothesized impacts. Using qualitative data to understand the results or reanalyzing the data together with other sources can lead the analyst to a richer causal model that makes better sense of the results. We discuss those strategies for building the evidence base below.

Use of Qualitative Findings

There is no easy way to validate a complete causal or theoretical model; a demonstration typically focuses on one possible cause and a small number of effects. To contextualize these findings, and to interpret where additional factors or unmodeled effects could be added to the model, qualitative data play an invaluable role. In particular, understanding the mechanisms by which an intervention produces an effect often comes from the story about why individuals react in a certain way that has no analog in the quantitative data. These stories might even appear in the text as narrative

interpretation of the impacts. Regardless, the exposition about mechanisms becomes much more plausible when based on interviews with participants who relate their actual perceptions and reasons for their reactions.

Many of the demonstrations SSA has conducted have included qualitative findings, based on focus groups, case studies, and detailed qualitative interviews. But these findings are not typically presented in the final report for a demonstration. For example, BOND included case studies and detailed interviews with participants but these do not appear in Gubits et al. (2018a/b), though those findings may inform some interpretations in the final report.

In general, quantitative findings are the sole focus in the final report. For example, Geyer et al. (2018, 60) reported that participants' "self-reported understanding of the benefit offset rules seems to have been influenced by their perceived need to understand them, their use of the offset, and related exposure to information." The BOND final report references such findings only obliquely, for example when discussing the low rates of correct understanding of program rules and the hypothesis that "one interpretation of these findings might be that most beneficiaries have no interest in working and thus pay little attention to how benefits would change with earnings" (Gubits et al. 2018a, 28). However, the BOND final report explicitly rejects that interpretation and privileges the quantitative data from the surveys: "we find no such evidence of differential understanding in Stage 1. In addition, Stage 2 treatment subjects who were working at baseline were not more likely to correctly understand the offset rules."

Similarly, Leiter, Wood, and Bell (1997) provide five anonymized narratives of participants' experiences in Project NetWork that provide a wide array of stories about the relative successes or failures of the interventions in that demonstration. The final report (Kornfeld et al. 1999) does not refer to that publication on the process results of the demonstration, nor its anonymized narratives, except as a citation in a footnote. However, the final report does draw heavily from the quantitative survey reports and various other prior analyses on services delivered. The narratives in the process report reinforce its conclusions that "substantial delays were encountered in obtaining diagnostic assessments" and that delays pushed back "provision of rehabilitation services" (Leiter, Wood, and Bell 1997, 47). For example, client profile 1 describes a client turned away by her state Vocational Rehabilitation agency but then connected to private sector agencies by Project NetWork and connecting to employment twice before a non-attorney representative advised her to exit the labor force and end her participation in Project NetWork. Client profile 2 describes a client turned away by her state Vocational Rehabilitation agency but getting help from a private vendor via Project NetWork.

To the extent that individual stories reflect the broader patterns measured in an impact evaluation, it would be valuable to include these stories in the final reports on a demonstration, to add a human element to the comparison of mean outcomes. Generally, though, our review of past SSA demonstrations indicates that qualitative

findings that appear in the intermediate or process reports do not appear explicitly in the final report. Instead, the patterns seen in qualitative findings earlier in a demonstration may inform the hypotheses about mechanisms and interpretation of findings that appear in the discussion sections of these reports.

Use of Data for Reanalysis and Longer-Term Outcomes

Pure replication of findings from a demonstration, "scientific replication" (meaning analysis using a different sample, a different population, or a somewhat different method), and additional analysis all add credibility to those findings and their contribution to the evidence base (Hammermesh 2007). Making a demonstration's data available to researchers and supporting their additional analysis is also important. A challenge for SSA's demonstrations is that their data are often restricted due to privacy laws, and the costs related to accessing and re-using the data have limited the number of pure replications and scientific replications conducted. Efforts to reduce those barriers to the data, as well as providing financial support to researchers to re-use them, could result in improvements to evidence-based policymaking.

Overcoming Data Barriers

One recent effort to reduce the costs to access the data is the inclusion in the DAF18 of additional demonstration information. The DAF18 includes a demonstrations and surveys extract that includes data on which SSDI beneficiaries and SSI recipients participated in one or more of the following SSA demonstrations and surveys:

- Accelerated Benefits (AB) demonstration;
- Benefits Entitlement Services Team (BEST) demonstration;
- Benefit Offset National Demonstration (BOND);
- Benefit Offset Pilot Demonstration (BOPD);
- Homeless Outreach Projects and Evaluation (HOPE) demonstration;
- Mental Health Treatment Study (MHTS);
- National Survey of SSI Children and Families (NSCF);
- Promoting Opportunity Demonstration (POD);
- Promoting Readiness of Minors in SSI (PROMISE) demonstration;
- Supported Employment Demonstration (SED); and
- Youth Transition Demonstration (YTD).

DAF18 also offers an NBS extract.

Though use of these DAF data is restricted to projects that meet the privacy and disclosure restrictions as disclosed to the participants in these data collections, the inclusion of the demonstrations and surveys extract in the DAF18 can reduce the data costs for replication and additional analysis.

Providing Financial Support

Another development is financial support for additional analysis of the demonstration projects through the Retirement and Disability Research Consortium, and through the Retirement Research Consortium and the Disability Research Consortium before that. The Retirement and Disability Research Consortium is an interdisciplinary extramural research program funded by the SSA through cooperative agreements with centers at Boston College, the National Bureau of Economic Research, the University of Michigan, and the University of Wisconsin. The solicitation for grant proposals encourages research employing a variety of approaches (e.g., descriptive and causal studies, simulations, etc.), using innovative methods, and drawing from new data sources (e.g., Occupational Requirements Survey data, data collected for demonstrations, etc.).

In addition to grant support, an opportunity to further reduce the costs of additional analysis of demonstration data is for SSA to develop new privacy and disclosure restrictions that make data from future demonstrations accessible for research purposes—of course, while ensuring the privacy of project participants. Indeed, this idea is reflected in the Commission on Evidence-Based Policymaking's final report (CEP 2017), as well as in provisions of the Foundations for Evidence-Based Policymaking Act of 2018 itself. Striking the right balance between access and privacy is a challenge, but if that balance can be found, then there is great potential to increase the return on investment from SSA's demonstrations. A useful step would be to re-examine the privacy and disclosure restrictions in prior demonstrations to identify potential changes toward improving access to data for research and reanalysis purposes. SSA has been engaged in this for years, but modifying systems of records is a very time-consuming process at best.

Synthesis of Findings across Demonstrations

Synthesizing findings across demonstrations can identify insights into program recruitment, enrollment, retention, efficacy, and effectiveness. The current volume tackles this challenge for recent SSA demonstrations related to disability policy. This effort should be ongoing, as new demonstrations add to the evidence base. These cross-demonstration insights can be useful for designing future demonstrations, implementing new programs, or informing an existing program or policy.

The process of recruiting and enrolling the target population into a demonstration project has proven to be challenging for some demonstrations. Ruiz-Quintanilla et al. (2006) summarized findings from the recruitment and enrollment process, as does Chapter 9. Notably, information is limited on the recruitment process for the demonstrations covered, as described in Chapter 9, despite the need for this type of information for planning future demonstrations. Ruiz-Quintanilla et al. report two key findings. First, the demonstration projects that target SSDI applicants instead of SSDI beneficiaries have relatively higher participation rates (between 14 percent and 22

percent, compared to around 5 percent for beneficiaries). Second, the strongest predictor of program participation is recent or current work experience. Chapter 9 indicates that these patterns continue to hold, though some subgroups may exceed our expectations based on average performance.

Finally, a broad assessment of the impacts across all of the demonstrations could identify opportunities to better target investments in future demonstrations. The assessment need not be limited to SSA, as other entities conduct demonstrations aimed at improving the employment and economic well-being of individuals with disabilities. For example, the US Department of Labor's Clearinghouse for Labor Evaluation and Research (known as CLEAR) identifies and summarizes many types of research, including descriptive statistical studies and outcome analyses, implementation studies, and causal impact studies. The Pathways to Work Evidence Clearinghouse and the related Employment Strategies for Low-Income Adults Evidence Review include a wide range of research assessing the effectiveness of the interventions reviewed. Commonly, a clearinghouse can provide a large set of individual studies without aggregating them appropriately to draw out their lessons.

With detailed information on implementation and service delivery in multiple sites and multiple demonstrations, it is tempting to look across a set of experiments to detect a pattern of where impacts are larger, then tell a story for why the impacts are larger in some situations and not others. To do so methodically, meta-analysis and meta-regression approaches hold promise. The fundamental idea of the meta-regression is that we have similar data on intervention impacts, and they vary systematically with features of the interventions that generate those impact estimates. When we combine all of the results in one regression, without picking and choosing any to support a story, we have adopted an approach that limits cherry-picking. Further, by weighting estimates according to their precision, we can get the most possible statistical power to answer the policy questions of interest.

SUMMARIZING THE LESSONS LEARNED ABOUT THE USE OF DEMONSTRATIONS

To date, SSA's demonstrations have mainly relied on rigorous, experimental evaluations that can answer causal questions convincingly. But those answers are valuable only if they yield actionable knowledge to inform policy and practice, even if that action is to not change the program in a direction hypothesized incorrectly to improve outcomes. This latter point is related to a key function of the Council of Economic Advisers (CEA) whose "analysis does have one important benefit, which is that it can help kill ideas that are completely logically inconsistent or wildly at variance with the data. This insight covers at least 90 percent of proposed economic policies" (Ben Bernanke, quoted in CEA [2016, 309]). Our read of the evidence in this volume is that employment services that include real-world work experiences have proven valuable in some past demonstrations and are being tried out in new populations in future demonstrations as a result, whereas benefit offsets have not produced the

desired effects, yet will be tested over and over again. That is, demonstrations have suggested productive activities to test anew, but have not killed off ideas whose time has come and gone.

We suggested earlier that past relevant demonstrations themselves might be subjected to a cost-benefit analysis before launching a new demonstration. To frame this cost-benefit analysis, we need to think broadly about the general goals of demonstrations ex ante, their added value, and how we might judge them ex post. Undertaking a new demonstration incurs substantial opportunity costs, and not just those related to government funds. We'd like to know that the benefit justifies the total cost. If a positive finding leads to the expansion of a policy, but a negative or null finding does not inform policy, then we should worry about the value of information or defects in the use of demonstrations.

A demonstration should address a specific policy-relevant question and generate answers that are useful for situations not yet observed. The data must also be sufficient to answer the question, and the findings communicated to be broadly understood. But broadly understood answers to policy-relevant questions are valuable only insofar as policymakers can use that evidence to formulate policy or program administrators can use that evidence to design and operate better interventions. This value is enhanced when demonstrations continue to produce evidence, which can happen when their data are combined with new data sources or reanalyzed in light of new developments.

The SSA has invested in planning and conducting major demonstrations. This volume is one of the first attempts to synthesize the lessons across demonstrations. Ongoing work should continue to situate new findings in this broader field of findings, and each new demonstration should be judged by how much actionable information it produces relative to the field of existing findings. This is a broader view of the value of a particular demonstration and does not relate to how we judge a contractor who faithfully executes a contracted demonstration with high quality. The ultimate value of a demonstration also rests on the value of the question being addressed and how the findings are used—a perfectly executed demonstration can still be a waste of time.

To facilitate understanding, a meta-analysis can be part of new demonstrations, as relevant, and the value of the information gleaned can be judged by the policy relevance of any shift in priors. An ex ante justification of a future demonstration can be based on a simulation of just such a meta-analysis, and a connection from possible estimates to policy actions implied by different true parameter ranges. Doing this would both clarify the goals of the demonstrations and make explicit the commitment of policymakers to use the information generated by them.

The clear first task of a demonstration is to answer the central research questions posed. But an additional important use for the results of demonstrations is to refine our theoretical understanding of causal relationships across interventions and individual behaviors. Understanding the role played by labor market features that are not under policymakers' control is important to interpreting findings. It argues both for replication at different times or under different conditions and for collecting qualitative

data that can motivate changes to the model. Our understanding of the strength of an effect could be interpreted quite differently if the effect is direct or if it is mediated entirely by another, much lower cost process. Alternatively, our understanding could be radically altered if the effect is large in the presence of some moderator, but nonexistent otherwise (or appears with the opposite sign).

The typical demonstration that tests one package of services versus a business-as-usual condition cannot address many crucial questions related to mediation or external validity of the findings, but a meta-analysis that incorporates planned variation across demonstrations of the right type can. Employing a cost-benefit lens not only for the intervention being tested but for the demonstration itself can point us to learning more. This process to learn more would first use past demonstrations to discover more than what appears in existing publications. Then it would design a next generation of demonstrations that illuminate the areas where we still find ourselves in the dark.

The best way forward maps answers to policy changes that are feasible and could produce large gains in well-being. Good use of demonstrations would maximize the expected return to a demonstration by generating evidence that changes policy to improve people's lives or prevents policy changes that would alter people's lives for the worse. That means making sure not just that the demonstration can produce an answer to the right policy question, but that the results can be used to design better policy.

Chapter 3

Comment

Jonah B. Gelbach
University of California, Berkeley

This chapter provides a wide-ranging assessment of how to make the most out of the Social Security Administration's (SSA) demonstrations. I found the chapter both comprehensive and insightful.

I focus my comments on one observation of Weathers and Nichols: "that past relevant demonstrations themselves might be subjected to a cost-benefit analysis before launching a new demonstration." They suggest such an analysis requires thinking "broadly about the general goals of demonstrations ex ante, their added value, and how we might judge them ex post." As Weathers and Nichols emphasize: "Undertaking a new demonstration incurs substantial opportunity costs, and not just those related to government funds. We'd like to know that the benefit justifies the total cost."

A good example is the chapter's discussion of program parameters' impact on the composition of program participants and applicants, i.e., entry effects. Weathers and Nichols discuss SSA's consideration of whether such effects could be productively studied for Social Security Disability Insurance (SSDI) using a traditional randomized control trial (RCT). Such a demonstration must target initial *non*participants, so SSA "would need to target a sample from the US population" as a whole. Reaching such a sample via a traditional demonstration would be expensive given the low population-level SSDI participation rate. An expert review suggested that a reasonable probability of detecting effects of interesting magnitudes would require something like nine million participants. Weathers and Nichols describe additional challenges the expert review raised; and in the BOND demonstration, SSA ultimately chose to focus on questions related to the reform's effects on current participants.

This discussion connects to a long-standing area of controversy among scholars studying causal effects of social programs: How much can we learn from RCTs, and should we concentrate our evaluation resources in that domain?

Larger-scale RCT social demonstrations have a long history, including the Negative Income Tax experiments of the 1970s and the Health Insurance Experiment of the 1970s and 1980s. When state-level welfare reforms were all the rage of the 1990s, numerous RCT demonstrations occurred.

The general argument for RCTs is familiar: They are supposed to balance differences in treatment and control groups, so that researchers and policymakers may be confident observed outcome differences are due to an intervention's causal effects rather than differences in confounders. For this reason, smaller-scale field RCTs have become more popular in recent years—especially in the area of development economics, for which economists Abhijit Banerjee, Esther Duflo, and Michael Kremer

won the 2019 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. Explaining its choice of topic area, the prize committee wrote that "the best way to draw precise conclusions about the true path from causes to effects is often to conduct a randomized control field trial" (Committee 2019, 5).

One frequently hears the term "gold standard" metaphorically applied to RCTs—and what could be better than gold! But there is a serious case that the actual gold standard contributed importantly to the scope of the Great Depression.[9] This is a rhetorical point to be sure, but it's a good reminder that apparently unassailable things can have their flaws. And the case against RCTs isn't just rhetorical. Nobel laureate in economics Angus Deaton and philosopher of science Nancy Cartwright wrote in a 2018 paper that "any special status for RCTs is unwarranted" (2). The good statistical properties of RCTs hold only on average, rather than in any particular RCT. And RCTs suffer precision-related issues, which help explain why estimating SSDI entry effects would have required so many participants. Another issue is treatment effect heterogeneity: some people will respond more than others, or even in the opposite direction, when facing changed policy parameters.[10]

The limits of RCTs related to treatment effect heterogeneity is a subject that has long been discussed in economics. Another Nobel laureate, James Heckman, pointed out in a 1995 paper with Jeffrey Smith that RCTs "do not identify the distribution of program gains unless additional assumptions are maintained." That is important because distributional considerations often are quite important to policymakers. To be sure, the fact that RCTs have their limits in the presence of heterogeneous effects doesn't mean distributional knowledge is out of reach. But it does mean some circumspection and careful attention to underlying theoretical considerations are warranted when considering RCT use. The 2019 Nobel prize committee itself embraced the role of economic theory in policy design (Committee 2019, 5).

At the risk of immodesty, I will point to my own work, co-authored with economists Marianne Bitler and Hillary Hoynes, using data from Connecticut's JOBS First welfare reform RCT demonstration project. In studying JOBS First, we were able to connect predictions from basic labor supply theory to the ways in which a change in program parameters could be expected to operate across the earnings distribution of demonstration subjects (Bitler, Gelbach, and Hoynes 2006). Our research was conducted entirely after the demonstration had finished, and it was possible only because MDRC's data were available for use by researchers via a not-too-onerous process. This raises an additional point: the value of making demonstration data publicly available for further study.

---

[9] Among other issues, adhering to the gold standard prevented central banks from using easier monetary policy to respond to negative shocks to demand. For an introductory-level discussion of central bank decisions, the gold standard, and the Great Depression, see Bernanke (2012). For a more extensive treatment, see Eichengreen (1996).

[10] Chapter 7 of this volume discusses that issue.

There are other criticisms of RCT demonstrations. For example, as Heckman and Smith pointed out in their 1995 paper, typical RCTs reveal only short-run policy reform effects. Of course, the same is true about many non-experimental evaluations. More generally, as Banerjee and Duflo (2009) note, the fact that RCTs have their problems doesn't mean that non-experimental approaches are immune to those same problems.[11]

What lessons can we draw from this discussion? First, well-designed, well-executed RCTs solve a particular class of statistical problem: they balance treatment and control groups on the distribution of confounding effects. That allows someone with the data in hand to estimate some kinds of parameters that may be of policy interest. But second, even perfectly implemented RCTs don't allow us to answer every question of interest—either because questions such as entry effects are by their nature difficult or expensive to study at all with RCTs, or because of the extent and nature of treatment effect heterogeneity.

The points above imply that whether RCTs are better than alternatives in any given context *depends*: It depends on the questions that are of interest, on the policy reform options, on the distribution of people's responses to policy reforms under consideration, and on what will be done with the information obtained from the demonstration.

Of course, whether RCTs are worth doing also depends on the alternative—we should always ask, "compared to what?"

There is a long history of non-experimental estimation in the social sciences. Both structural and reduced form econometric methods have developed in important part for the purpose of answering the kinds of questions that RCTs would answer if they existed. These points are not unknown to the discussion of SSA demonstrations, as the discussion of entry effects that Weathers and Nichols offer illustrates. They cite an SSA-funded RAND paper by Nicole Maestas, Kathleen J. Mullen, and Gema Zamarro (2010), titled *Research Designs for Estimating Induced Entry into the SSDI Program Resulting from a Benefit Offset*, which describes two RCT alternatives—stated preferences and structural estimation using variation from past policy changes. These authors considered but rejected alternative approaches, including more complex structural models.

I suggest here that even where RCTs are feasible to design and administer at manageable cost, it is not obvious that they are always the best choice. One way to look at this issue is to recognize that the choice to use an RCT to study a question is itself a policy choice. The internal logic of the contention that RCTs are necessary for better policy study therefore requires randomizing whether RCTs are used to study

---

[11] "[A] although some of these issues are specific to experiments…, most of these concerns (external validity, the difference between partial equilibrium and market equilibrium effects, nonidentification of distribution of effect) are common to all microevaluations, both with experimental and nonexperimental methods" (2009, 159).

questions. That seems unlikely. What we have available is the considered ex ante judgment of experts. SSA ought to use that resource liberally.

I have one final suggestion.

The federal government ought to invest in making SSA's data more available to researchers operating outside either the agency itself or its contracted parties. There are lots of highly skilled researchers who want to study questions that are or would be of interest to policymakers but who aren't able to do so because they can't get data. The federal government could radically increase the amount of available research knowledge by making existing SSA administrative data more publicly accessible. Of course there are privacy considerations, and program operations must continue without interruption. But perhaps the federal government should consider whether the next demonstration is likely to lead to information as valuable as might be gained were it to spend some of its resources figuring out how to productively share data for wider study.

In sum, I applaud Weathers and Nichols's general suggestion that substantial thought should be given to whether particular demonstrations are worth the expense and time it will take to conduct them. There are alternatives, including non-experimental study in particular settings and wider data access in general. It is to SSA's credit that the agency has commissioned this volume, and it will be to all of our benefit if the agency follows these authors' suggestion.

Chapter 3

Comment

Elizabeth H. Curda
US Government Accountability Office[12]

Over the last 20 years, the Social Security Administration (SSA) has carried out many demonstration projects and spent hundreds of millions of dollars doing so. Given the time and money invested in them, demonstration projects need to be carefully designed so that the results will inform important improvements to outcomes for Social Security Disability Insurance beneficiaries, Supplemental Security Income recipients, and taxpayers.

Chapter 3, "Improving the Use of Demonstrations," suggests an array of promising practices to enhance the effectiveness of SSA demonstration projects. These suggestions fall into three main categories: methodological, process, and communication. Many of these promising practices dovetail with prior Government Accountability Office (GAO) analyses and recommendations. Though we recognize that SSA has implemented many of these recommendations, it is worthwhile to highlight where GAO has taken a similar position. More broadly, portions of Chapter 3 echo best practices for project management, as well as internal control standards that apply to all federal programs. The following paragraphs highlight GAO findings, recommendations, standards, and best practices that add further impetus to the authors' recommendations to improve the use of demonstration projects.

METHODS

Of the many important points made in Chapter 3 relating to the effective design of demonstration projects, one that stands out is the suggestion to employ falsifiable logic models in order to better identify programs that are ready for rigorous outcome assessments. GAO has long recommended the use of logic models in developing programs and evaluations of those programs and greater use of the falsifiable logic model has the potential to ensure a demonstration program's process has been sufficiently vetted and improved prior to employing more costly and consequential outcome evaluations (see e.g., GAO 2002).

The authors also suggest that researchers can and should leverage demonstration projects to test multiple intervention options and causal channels through multistage, multi-arm, or factorial design of interventions and experimental evaluations. Doing so could be a good way to increase the return on investment of a given demonstration.

---

[12] The views expressed in this comment are those of the author and do not necessarily represent the views of the Government Accountability Office or the US federal government. I am grateful to Jessica Rider, a Senior Economist at the GAO, for her assistance in drafting preliminary versions of these comments.

However, doing so also increases complexity and requires careful design to be effective. GAO's work on designing evaluations stresses the need to be clear about the evaluation questions at each phase of project and what is to be assessed—process versus outcomes versus impact of alternative interventions, for example—and to select appropriate measures and criteria for success at each stage of a program's implementation (e.g., program uptake among eligible individuals, changes in key program outcomes, use of program resources) (GAO 2012a).

GAO has emphasized the necessity of assessing project effects compared to a counterfactual of no intervention. For instance, in GAO's 2008 report on SSA's demonstration projects, GAO found that some demonstrations at that time did not assess the project's effects compared to what would have happened in its absence. GAO also found that planning for the evaluation has to be part of the demonstration project's design. As part of that report, GAO recommended that SSA implement clear written policies and procedures that are consistent with standard research processes and federal internal controls standards. As a result, SSA developed a Demonstration Project Guidebook, which outlines the agency's policies, procedures, and mechanisms for managing and operating its demonstration projects. This Guidebook could serve as the repository for any key insights and best practices SSA adopts from these lessons learned.

A key aspect of internal control is identifying potential risks to the success of a demonstration project in advance and being prepared to analyze and respond to the risks. This principle encompasses, at a high level, some of the methodological practices the authors highlight in Chapter 3, such as identifying and documenting tradeoffs in scoping the project, identifying ways that the proposed methods or timeframes may fail to meet the needs of policymakers, and understanding potential sources of bias in the analysis.

PROCESS

In Chapter 3, the authors highlight the need to build a better evidence base by, among other things, using qualitative information to provide context and explore causal mechanisms. Taking that a step further, considering participant voices in the planning and design of an intervention often yields new insights about potential risks to agency actions. For example, in a 2010 forum held by GAO, stakeholders, including those with a participant perspective, noted that new SSA disability benefits, services, and programs need to be carefully structured to avoid unintended consequences and that the costs and benefits to participants must be considered in program design.

Another process improvement is to take steps to ensure transparency about changes to the demonstration along the way. In a recent report on Medicaid evaluations, GAO found that changes during a demonstration can cause problems and affect the quality of the evaluation—changes to the design of the demonstration, the sample, related policies that may affect participants, etc. should be documented along with plans for how those changes will be handled in the evaluation (GAO 2019).

COMMUNICATION

Chapter 3 stresses the need to involve stakeholders early and often, as well as to disseminate interim results to key stakeholders. This is a critical best practice and while this type of collaborative process takes time, it typically leads to less rework and more robust results.

The authors also state that demonstration findings should be leveraged to affect policy through good communication. This is critical, but not always practiced. GAO has recommended as recently as 2018 to the US Department of Health and Human Services that it provide rigorous final evaluation reports and publicly release the findings of demonstration projects (see e.g., GAO 2018).

And finally, going beyond the focus on individual demonstration projects, GAO has previously identified more than 40 programs managed by nine different agencies that provide a patchwork of employment support for people with disabilities. We reported in 2012 that these programs lacked a unified vision, strategy, or set of goals to guide their outcomes. GAO has recommended since 2012 that the Office of Management and Budget work with federal agencies to coordinate the development of a set of unifying, government-wide goals for employment of people with disabilities (GAO 2012b). Such an effort could provide much needed focus and impetus for designing demonstration projects that align with federal employment goals. It could also help agencies take greater advantage of the wealth of data collected by different federal agencies, which currently requires herculean efforts by agencies and researchers to obtain and use in these important demonstration evaluations.

Chapter 4

The Return to Work in Disability Programs: What Has Been Learned and Next Steps

Jesse Gregory
University of Wisconsin
Robert A. Moffitt
Johns Hopkins University

The United States has two major government programs for individuals with disabilities. The first, the Social Security Disability Insurance (SSDI) program, provides cash benefits and health insurance in the form of Medicare to individuals younger than age 66 or 67 who meet a test for severe disability and who have sufficient past earnings in jobs covered by the Social Security system.[1] The amount of the monthly cash benefit is determined by the level of past earnings. The second, the Supplemental Security Income (SSI) program, is means-tested and provides cash payments and Medicaid coverage to individuals who meet the same disability test as in SSDI, but who have low income and assets and hence qualify under a current means test rather than under a work-history test.[2] The basic monthly benefit level depends on marital status and living arrangements, unearned income, and other factors, but not on past work history.[3]

Both SSI and SSDI are intended to provide support only to those who have no or little capacity to work, and the history of the discussions even at the time of creation of these programs reveals a concern with work disincentives (Berkowitz 2013, 2020). An emphasis on rehabilitation has always been a central focus of the programs. However, as the caseloads of both programs began to rise in the 1980s and accelerated in the 1990s and 2000s (but have recently peaked and begun to fall), discussions of how to promote returns to work among SSDI beneficiaries and SSI recipients increased. This concern led to considerable discussion and congressional attention and academic research but also led the Social Security Administration (SSA) to conduct a number of demonstrations, almost always using randomized control trial methods, to test possible reforms to the programs to increase work, employment, and earnings.

---

1 SSDI is a social insurance program, where the premiums are the tax contributions made from covered earnings. We are referring to worker beneficiaries, with which this chapter is concerned, not disabled adult children or disabled widows or widowers.

2 We will only be concerned with the SSI program for adults, not children or the elderly. Elderly and child SSI recipients have different eligibility rules.

3 It is also possible to receive benefits from both programs (these are called "concurrent beneficiaries"). The programs both have work incentives and access to Vocational Rehabilitation and the Ticket to Work programs that we discuss below.

Though both SSDI and SSI intend to provide benefits only to those with severe disabilities and not partial disabilities, some concurrent beneficiaries have some ability to work—residual functional work capacity (often called residual work capacity or "work capacity" for short)—although the number with the capacity for substantial work levels is uncertain. There is widespread agreement that those with capacity to work should be encouraged to use that capacity through labor market activity, and that those with substantial capacity who can leave the program should be encouraged to do so.[4]

One major concern about the structure of the current SSDI program is that the rules of the program governing work might provide meaningful work disincentives to those with such residual capacity. The rules follow from the philosophy of the program, which is to divide the population into those who can work and those who cannot and to provide benefits only to the latter—to provide a safety net for those unable to work at a minimally self-sustaining level (Substantial Gainful Activity, or SGA). Several steps in the determination process are designed to make that division of the applicant pool into the two groups, but there are inevitably errors in that determination. An individual's work capacity can change over time, as well. Once in the program, an individual is designated to be able to work if they demonstrate that they do, in fact, have residual work capacity through actual work and earnings or if medical determinations show that they no longer meet the disability standard for the program. Therefore, the program rules are designed to test whether an individual's disability prevents them from performing work considered SGA.[5]

The SSI program provides less direct financial work disincentive than does SSDI, imposing a 50 percent benefit reduction rate (BRR) on earnings (after a disregard and deductions) rather using earnings per se as a test of disability (although both provide work programs of other types and special incentives for the blind). However, few recipients of SSI work at high levels, similar to the SSDI program, and the desire to increase employment and labor force engagement among those with residual work capacity is similar to that for SSDI beneficiaries.

In the first section of this chapter, we review the rules of the SSDI and SSI programs with a focus on those governing work and earnings. The second section discusses conceptual issues that should influence thinking about how those rules might affect individual motivations to work. The third section of this chapter reviews the evidence from a number of SSA demonstrations. The demonstrations reviewed are only those designed for individuals who are already beneficiaries or recipients of one of the two programs or have applied, and not demonstrations that attempt early interventions prior to application (those are covered by Hollenbeck [2021]). The fourth

---

[4] Both programs explicitly state that encouraging beneficiaries and recipients to work if they can is one of the goals of the programs (SSA 2020b, 6; SSA 2020h, 1).

[5] There have been many proposals for expanding the SSDI program, or creating new ones, to cover the partially disabled. See Maestas (2019) for one such recent proposal. We consider such proposals briefly in the final section of this chapter.

section draws some general lessons learned from the demonstrations. The fifth section suggests directions for new demonstrations and demonstration designs that might be considered. A short summary concludes the chapter.

STRUCTURE OF THE SSDI AND SSI PROGRAMS

Social Security Disability Insurance Program (SSDI)

The SSDI program, begun in 1956, had slightly more than 8 million disabled workers in 2019 (SSA 2020b).[6] About 30 percent establish eligibility on the basis of an intellectual disability or other mental disorder, 34 percent on the basis of a musculoskeletal system or connective tissue disability, and the rest on the basis of a range of other impairments. The caseload has grown for several decades, with disproportionate growth among those with mental and musculoskeletal impairments.[7] To the extent that residual work capacity differs by impairment type, these trends could affect aggregate caseload residual capacity. The caseload peaked in 2014 and has declined since, partly from demographic trends and partly from administrative changes (Maestas 2019).

There is a large literature on the sources of the rapid caseload growth, most of which is outside the scope of our review. However, one suggested contributor to growth is the long-term decline in the demand for unskilled labor in general—a trend at least since the 1980s—with a consequent decline in average wages paid in the private sector for low-skilled jobs. Autor and Duggan (2000) argue that the decline in those wages has been one of the forces pushing up caseloads, as individuals with disabilities tend to work at such jobs. If so, the implication is that increasing employment of SSDI beneficiaries or increasing the rate of exit from SSDI to employment could have become more difficult.

Exits from SSDI because of a successful return to work are rare. Though almost 10 percent of beneficiaries have had their benefits terminated in recent years, most of these are the result of death or reaching the full retirement age. Only 0.7 percent of SSDI worker beneficiaries terminated benefits from a successful return to work in 2019 (SSA 2020b, Table 50).[8] Those with mental impairments had higher rates of termination whereas those with musculoskeletal impairments had lower rates. Some beneficiaries, as described below, have benefits withheld for a month because of high earnings but are not terminated. Another 0.7 percent of beneficiaries experienced such

---

[6] This constitutes 89 percent of the caseload, the rest consisting of disabled adult children and disabled widows and widowers.

[7] The growth in the share of SSDI caseload with mental and musculoskeletal impairments corresponds with an increase in the prevalence of these impairments among the broader US population with disabilities (Mojtabai 2011; Freburger et al. 2009).

[8] Longitudinal studies show higher rates of exit (Liu and Stapleton 2011).

withholding in December 2019. Those with mental impairments had about the same rate whereas those with musculoskeletal disorders had a lower rate.

Application for SSDI involves a five-step process that begins with a determination of whether the applicant is engaged in SGA, which in 2021 was $1,310 per month for a nonblind applicant and $2,190 for a blind applicant. After also determining whether the disability is severe, the individual's disability must be found in a list of impairments. Then an assessment is made of whether the individual can return to their pre-disability job or to any other job in the economy. Denials can be appealed, and the ultimate decision in individual cases can sometimes take several years. There have been many proposals to reform elements of this process, but most are not directly related to work disincentives other than the SGA criterion itself (we mention an exception in the next section).

After an SSDI award, a number of work-related rules are present in the program. The first, called the Trial Work Period (TWP), begins if earnings are above a threshold, allowing the beneficiary to experience no reduction in benefits for up to nine months within a 60-month window.[9] Those who complete a TWP have a 3-month Grace Period the first month the beneficiary works SGA, followed by an EPE, which is a 36-month reentitlement period during which benefits are withheld in months with earnings at or above the SGA level but fully paid in months with earnings below the SGA level. After the 36-month period elapses, any subsequent month with earnings above the SGA level results in benefit termination and exit from the program (but with extended Medicare coverage). An expedited reinstatement application is allowed if the individual becomes unable to work again within the next 60 months. In the 2014–2018 period, about 2 to 3 percent of beneficiaries per year used the TWP and about half of those completed a TWP (SSA 2020a, Tables 1 and 58).[10]

There are a large number of other work-related SSDI rules that have not been directly studied (i.e., have not been the subject of randomized trials). These include rules governing unsuccessful work attempts, deductions for impairment-related work expenses, continued payments under certain circumstances for those engaged in Vocational Rehabilitation (VR), provisions for extended Medicare coverage after completion of the TWP, options to buy Medicare for some period after benefit termination, and a variety of services intended to help those who wish to return to work.[11]

---

[9] There are a number of other calculations of earnings amounts prior to application of the TWP rule. See the *Red Book* (SSA 2020e) at https://www.ssa.gov/redbook/.

[10] Again, the rates could be higher in longitudinal statistics.

[11] These other rules are discussed in SSA's *Red Book* (https://www.ssa.gov/redbook/eng/resources-supports.htm).

Supplemental Security Income Program (SSI)

Enacted by Congress in 1972, the SSI program replaced a state-level system of support for the low-income, aged, blind, and disabled population with a federal program with nationwide benefits and eligibility criteria. The definition of disability used by the SSI program for initial award consideration is the same as that for the SSDI program, but SSI also requires that income and assets fall below certain specific levels, as well as having a number of citizenship requirements.[12]

After an award, SGA does not come into play because under SSI there is no direct post-award earnings test for being disabled as there is for SSDI.[13] SSI instead operates much more like a traditional transfer program, as it is intended to do, with a maximum payment made to those with zero earnings and with a reduced benefit as earnings rise (after certain deductions). The BRR is 50 percent, not far from the rate in many other means-tested transfer programs. There are also a number of other services designed to encourage a return to work, some shared with SSDI and some only for SSI.[14] Unearned income (with certain exceptions) reduces benefits dollar-for-dollar, following from the philosophy of the program as means-tested rather than as social insurance. This treatment allows SSI recipients to receive SSDI (termed concurrent beneficiaries) as long their SSDI benefit is not so large as to make them ineligible for an SSI payment after disregards. Finally, a number of states also have supplemental SSI programs.[15]

The percentage of blind and disabled SSI recipients ages 18 to 64 who work was 7.9 percent in 2007 and fell to 7.0 percent in 2019[16]; 2019 employment rates are higher for those with autistic disorders (17 percent) and intellectual disorders (12 percent), but are close to zero for many other diagnostic groups. Employment rates also are highest among recipients in their 20s and 30s. Exit rates because of work are low: of those receiving benefits in 2018, only 1.4 percent had achieved high enough earnings in 2019 to no longer be eligible for the program.[17]

---

12 SSI and SSDI have different look-back periods for back benefits after an award is made. The earnings limits increase with inflation, but the asset limits do not and have been held constant in nominal terms for many years. Those admitted to SSI are generally eligible for Medicaid on the basis of their income.

13 What are termed "Section 1619 waivers" allow recipients to work above the SGA level.

14 Refer to the *Red Book* (SSA 2020e) at https://www.ssa.gov/redbook/eng/main.htm.

15 See Duggan, Kearney, and Rennane (2016) for a description of the range of state programs.

16 See SSA (2020h, Table 42) and earlier years back to 2007. Figures earlier than 2007, but for work rates for all SSI recipients regardless of age, peaked in 2000 and have been falling since then (SSA 2020h, Table 40).

17 This is for recipients in 1619(b) status. See SSA (2020h, Table 44: 102,867 divided by 6,971,113). This earnings level is generally above the SGA level. For example, in 2020 an individual living alone with no deductions other than earnings and with no income other than earnings could earn up to $1,436 in a month before losing eligibility (the SGA threshold is $1,220). Longitudinal statistics could show different exit rates.

CONCEPTUAL ISSUES

In this section, we discuss conceptual issues surrounding providing work support and work incentives to recipients of the SSI program and beneficiaries of the SSDI program and to the severely disabled population in general. We draw upon several decades of thinking and research on this topic from the economic literature on work incentives of transfer programs for the needy in general, and where those apply to recipients of disability benefits, as well as from the research specifically on SSDI and SSI. There are many general lessons from this literature that should help guide thinking for these programs, but also a number of dimensions that differ for the SSDI and SSI programs.[18]

The conventional view of employment decisions, and of decisions about working part-time, full-time, or some other hours of work, is that individuals compare the value of working at different levels to the value of not working. The value of working includes financial compensation, fringe benefits, and health insurance provision, for example. And most individuals also place non-economic value on working because it improves an individual's feeling of self-worth and contributing to society. The value of not working includes the value of income available in that status, as well as the value to the individual of the reduction of time spent at work (or more generally, the disutility of work). Health considerations always play a role in the disutility of work, and that is an obvious element integral to labor force decisions of those with disabilities. Transportation costs and other fixed costs of working also play a role in the value of working, as does the availability of accommodation. The availability of income and care support from others in the household is important to the value of not working.

Lack of knowledge of the labor market for individuals with disabilities could also play a role insofar as individuals who do not work may not know the types of jobs that they might have open to them, and it may take substantial time and effort to search for jobs to acquire that information. On the labor demand side, the number of job openings that are suitable to the individual is important, as is the state of the business cycle.[19] There are also dynamic considerations, the two most important being the positive impact of working on future earning power (human capital)—most important for younger workers—and the negative impact of approaching retirement on the incentive to invest in acquiring new skills—a consideration important for older workers.[20]

All these considerations can be safely assumed to be part of the benefits calculus of individuals with disabilities. Medical conditions increase the disutility of working, and fixed costs of work as well as transportation costs are likely to be important, as is

---

[18] See Moffitt (2016) for reviews of the major US means-tested transfers and their effects on work effort.

[19] Discrimination against individuals with disabilities may affect the jobs available, and racial discrimination may also hinder job opportunities for those with disabilities.

[20] There may be other dynamic considerations.

lack of information about jobs that are suitable and provide some degree of accommodation. These factors are likely to be more important for individuals with disabilities than for individuals out them, and all will tend to discourage work. As for SSDI and SSI, if an individual is not working, those programs provide cash support and health insurance, both of which also generate work disincentives by providing support to those who do not work A major problem with assessing the effects of SSDI and SSI rules on labor force decisions of their beneficiaries and recipients is that the general responsiveness to the factors affecting those decisions is not known from the research literature. The basic building blocks of the economic model outlined above are the concepts of substitution and income effects. The "substitution effect" refers to how much more individuals will work if the financial gain from working increases, and the "income effect" refers to how much less individuals will work if they are given higher benefits for not working. When the "dynamic" considerations mentioned above are considered, the concept of discount rates (how individuals weigh future outcomes), and the returns to human capital from the types of jobs that individuals with disabilities have open to them, are relevant to individual decisions. Although substantial existing evidence estimates these factors for those without disabilities, estimates are largely absent from the research literature for those with disabilities.[21] However, many of the demonstrations we will review in the next section have suggestive evidence on some of these factors.

Returning to the general research literature on work incentives in transfer programs, most economists view the problem as one that requires finding a balance between providing support for a needy population in the form of cash, medical care, or subsidies of various kinds, and the preservation of at least some incentives to work. The goal of the programs is to continue to provide support to those who need it, while encouraging and assisting those with residual work capacity to use that capacity. The role of research in this enterprise is to try to identify mechanisms that reduce the magnitude of this central tradeoff by finding reforms that make it easier to achieve that balance.

This tradeoff is handled differently in SSDI and SSI. In SSDI, beneficiaries ultimately deemed capable of working at the SGA level are regarded as not disabled and experience benefit termination. But in SSI, working above the SGA level is allowed; recipients with higher levels of earnings are considered eligible, as long as they are still judged to meet the medical definition of disability and meet the countable income test. In SSDI, beneficiaries who consistently work above the SGA level are only eligible for a limited time.

When thinking about ways to manage the tradeoff, a well understood issue in the existing literature is that transfer programs with high BRRs provide work disincentives to recipients, but lowering those rates does not by itself lead to program exit. Indeed,

---

[21] Autor and Duggan (2007) argue that income effects for those with disabilities can be estimated from the Department of Veterans Affairs program.

the original Negative Income Tax proposals by Milton Friedman were intended to allow recipients to keep receiving benefits after beginning to work, and that this supplementation would be permanent, not temporary. The SSI program accepts that tradeoff to a greater degree than the SSDI program does.[22] The evidence we will review in the next section on SSDI suggests that allowing beneficiaries to work above the SGA level, but only temporarily (because, eventually, they will be terminated from the program), might not have a very large impact on labor force engagement.

Another principle from the literature on low BRRs is that reducing those rates has ambiguous effects on labor supply because, by extending upward the range of earnings over which benefits can be received, the BRR is increased for some individuals (Hoynes and Moffitt 1999; Moffitt 1992b).[23] For individuals with earnings that fall beyond the initial phase-out region but fall within the expanded phase-out region, a substitution effect will tend to reduce labor supply. Additionally, an income effect will tend to reduce labor supply for all individuals whose benefits increase at their initial earnings level. This has a direct application to benefit cliffs such as that in the SSDI program for the SGA level. Everyone dislikes benefit cliffs, but smoothing them out has ambiguous effects on average levels of labor supply: Increases in benefits tend to reduce labor supply through income effects, and some individuals face lower BRRs whereas others face higher rates, all with an ambiguous sign on the net effect. Smoothing out benefit cliffs implies increasing some beneficiaries' work levels and reducing them for others, which implies a tradeoff across individuals—we must consider whether a higher-earnings group that gets the smoothing is more worthy of support than the group not subject to the smoothing.[24]

These simple lessons from other programs take a different form in SSDI, with its work rules in the form of the TWP and the EPE. The TWP imposes a zero BRR, the EPE reentitlement period imposes a benefit cliff for earnings at the SGA level, and benefits are terminated for earnings above the SGA level in the post-reentitlement EPE period (ignoring the Grace Period). Altering the TWP by imposing a nonzero BRR above the TWP threshold or smoothing the benefit cliff during the reentitlement period would both have ambiguous net effects on work levels. But, unlike means-tested programs where high enough levels of work result in fairly immediate termination from the program, in SSDI the beneficiary has up to eight years before termination occurs (60 months for the TWP but which can occur in any window after receipt begins, and 36 months for the EPE reentitlement period). A beneficiary therefore has quite a long time to establish skills and work patterns before having their benefits terminated. Probably also important is the uncertainty that individuals have of their own earning ability and how difficult it will be to sustain working if their medical

---

[22] Administrative costs in conducting large numbers of work continuing disability reviews (CDRs) in SSDI for very small supplements could play a role here.

[23] Levy (1979) was the first paper to note this result in the welfare program literature.

[24] However, lowering BRRs often unambiguously increases the probability of being employed at all.

conditions remain as barriers, although the purpose of the TWP and EPE is to allow the beneficiary to test out their ability to work. The usual presumption is that, in the face of uncertainty, most individuals are risk averse and are less likely to undertake actions such as working in the face of that uncertainty. These considerations are absent in SSI, which does not have an earnings test per se to determine whether disability continues.

A consideration sometimes mentioned in the literature on other transfer programs is that work can increase human capital and therefore earning power in the labor market, which could mean that short-term incentives to work could have a longer-run impact by increasing exit rates from the program. However, most unskilled jobs have relatively low rates of return to human capital and so any effect of this kind is likely to be small.[25]

Another "dynamic" consideration results from the unique feature of SSDI and SSI (not present in other transfer programs) represented by their extensive application process, with its determination of medical condition and work capacity. That the determination period can be quite long for some applicants initially denied (although not the majority), and that individuals often do not risk working even below the SGA level during that period, could reduce long-term employability and even the desire for work (Autor et al. 2015). It is therefore possible that the features of the application process have effects on labor force engagement among SSDI beneficiaries and SSI recipients quite apart from the effects of BRRs, VR, and the other central work supports provided to recipients after an award.

REVIEW OF SSA DEMONSTRATIONS

The changes in policies, programs, and services to promote SSDI beneficiaries' and SSI recipients' return to work that SSA demonstrations have studied fall in four broad categories: (1) those that operate through changes to financial incentives, (2) those that involve modifications in the use of VR services or other direct employment services, (3) those that offer new employment programs specifically aimed to individuals with mental impairments, and (4) those that modify health insurance coverage. We divide our survey of SSA demonstrations based on this categorization, but recognizing that some demonstrations involve multiple interventions to some degree.

[25] And, in fact, encouraging employment can reduce future human capital to the extent it reduces time that would otherwise be spent in education or formal training, as noted by Heckman, Lochner, and Cossa (2003). The discouraging effect on education is actually one reason why the Ticket to Work program doesn't include all those of working age (i.e., age 16 and older). However, education and training could be less important for the older SSDI population.

Exhibit 4.1. Demonstrations Reviewed

Name	Status	Years	Population	Intervention	Results to Date
			Financial Incentives		
Benefit Offset National Demonstration (BOND)	Completed	2011–2015[a]	• SSDI beneficiaries • Concurrent beneficiaries	$1 for $2 benefit offset for annual earnings	Increased benefits, differently signed effects on earnings: above BYA (+), above 2 × BYA (−)
Promoting Opportunity Demonstration (POD)	In progress	2018–	• SSDI beneficiaries	Elimination of TWP/EPE, $1 for $2 benefit offset	Results not yet available
State Partnership Initiative (SPI)	Completed	1999–2004	• SSDI beneficiaries • SSI recipients • Others with disabilities	Combinations of case management, job training, employer outreach, and other employment services	No experimental results to report due to lack of credible comparison groups
State Partnership Initiative (SPI) SSI Work Incentives Demonstration Project	Completed	2001–2004	• SSI recipients • Concurrent beneficiaries	$3 for $4 earnings exemption, allowed asset accumulation, suspended certain CDRs	No experimental results to report due to lack of credible comparison groups
			Vocational Rehabilitation		
Project NetWork	Completed	1992–1995	• SSDI beneficiaries • SSI applicants and recipients	Case management services	Increased service receipt, small short-run increases in earnings
Ticket to Work (TTW)[b]	Completed	2002–	• SSDI beneficiaries • SSI recipients • Concurrent beneficiaries	Employment services, financial incentive to service providers	Increased service receipt, entry of services providers
Ohio Direct Referral Demonstration (ODRD)	In progress	2020–	• Young SSDI and SSI applicants and beneficiaries/recipients	Direct referral to VR	Results not yet available

Name	Status	Years	Population	Intervention	Results to Date
			Mental Impairments		
Transitional Employment Training Demonstration (TETD)	Completed	1985–1987	• SSI recipients with intellectual disability	Job placements, specialized OJT, postplacement support	Increased employment and earnings
Mental Health Treatment Study (MHTS)	Completed	2006–2010	• SSDI and concurrent beneficiaries with schizophrenia or affective disorder	Comprehensive set of services including IPS	Increased employment and earnings
Supported Employment Demonstration (SED)	In progress	2017–	• SSDI and SSI denied applicants with mental impairments	1. Basic services including IPS 2. Plus services of a nurse coordinator	Results not yet available
			Health Insurance		
Accelerated Benefit (AB)	Completed	2007–-2010	• SSDI beneficiaries in 24-month Medicare waiting period	1. Health Insurance coverage 2. Plus other services	Increased employment in 2nd year, none by 3rd year

Key: BYA=BOND Yearly Amount. TWP=Trial Work Period. EPE=Extended Period of Eligibility. CDR=continuing disability review. IPS=Individual Placement and Support. OJT=on-the-job training. VR=Vocational Rehabilitation.

[a] Some beneficiaries will continue to receive benefits into 2022.

[b] Ticket to Work was not an SSA demonstration. We include it in this review because it is an important component of recent policy facilitating return to work and because its intentionally staggered rollout provides quasi-experimental policy variation that was the focus of evaluations of the program.

Demonstrations Focused on Financial Incentives

Benefit Offset National Demonstration (BOND)

The largest demonstration to date studying the impact of removing financial work disincentives on return-to-work outcomes was BOND, which was mandated by Congress in the Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act).[26] Beginning in 2011, BOND used random assignment to select treatment and control groups from a large nationally representative cross section of the SSDI beneficiaries under age 60. All SSDI beneficiaries (including those concurrent with SSI) between ages 20 and 59 from sampled Area Offices and not in another SSA demonstration were included in the main BOND sample, which was called Stage 1 (*N*=77,101 treatment group, *N*=891,429 control group). Different from many

[26] Because of its importance, we spend more time reviewing it than the others that follow.

demonstrations, this meant that the BOND Stage 1 sample was not limited to volunteers. A smaller sample of SSDI-only beneficiaries predicted to be most likely to use the offset were recruited to volunteer for a second intervention, called Stage 2 (treatment group *N*=7,895, control group *N*=4,849). All participants in Stage 2 were volunteers. Treatment or control status was randomly assigned to participants in both stages.[27]

The treatment groups in both Stage 1 and Stage 2 received a modification to the normal SSDI benefit formula that removed the cash cliff faced by SSDI beneficiaries who have completed a TWP. Under the normal SSDI program rules, following the TWP, monthly earnings above the SGA level result in the loss of SSDI benefits. BOND did not alter the rules governing the TWP itself. Instead, under the modified rules for the BOND treatment groups, following the TWP, earnings above a related earnings threshold for annual earnings called the BOND Yearly Amount (BYA) resulted in a $1 reduction in benefits for each $2 of additional earnings. This modified schedule lasted for five years, referred to as the BOND participation period. The treatment group also received work incentives counseling on par with what is available through the Work Incentives Planning and Assistance program under current law. The Stage 2 treatment group was divided into two arms. Both treatment arms received the $1 for $2 benefit offset. Additionally, one treatment arm received enhanced work incentives counseling and the other standard counseling. The goal of Stage 2 was to learn about differences in the effectiveness of passive versus active counseling, as the enhanced counseling involved staff reaching out to beneficiaries unsolicited.

The change to an annual accounting system added an additional layer of administrative complexity. Each year, beneficiaries who intended to work were asked to submit an Annual Earnings Estimate (AEE).[28] If anticipated earnings exceeded the BYA, monthly benefits were reduced using the $1 for $2 offset calculation applied to the AEE. If actual annual earnings differed from the AEE, SSA made an adjustment either by paying additional back benefits or recouping overpayments.

The final BOND evaluation report focused on total earnings and total SSDI benefits between 2011 and 2015 as the confirmatory outcomes of interest. Relative to the control group, the Stage 1 (nationally representative) treatment group received average total benefits over the five-year window that exceeded those of the control group by $655 (relative to a control mean of $53,490). The mean earnings of the treatment group were nearly identical to that of the control group.[29]

---

[27] See the final evaluation report by Gubits et al. (2018a/b) for additional details on BOND.

[28] This process was similar to the one used to apply the Social Security retirement earnings test.

[29] Though employment per se was not the major focus of the BOND reports, the Stage 1 treatment group had a statistically significant small increase in employment (one-third of a percentage point impact on the probability of any employment 2011–2015 [control group mean was 22.1 percent] and on the number of years of employment [+.01 of a year, control group mean was 0.67 years]).

Though mean earnings were similar for the Stage 1 treatment and control groups, the distribution of earnings differed in several notable ways that suggest that BOND had opposite-signed impacts on earnings in different parts of the earnings distribution.[30] A larger fraction of the treatment group than the control group had annual earnings above BYA (+0.23 percentage points from a 2014/2015 base of 2.8 percent). However, a smaller fraction of the treatment group had annual earnings above 200 percent of the BYA (−0.1 percentage points from a 2014/2015 base of 1.4 percent) or above 300 percent of BYA (−0.08 percentage points from a 2014/2015 base of 0.76 percent). The small magnitude of BOND's impacts on the share with earnings above BYA, 200 percent BYA, and 300 percent BYA in part reflects the small size of the subgroup that was directly affected by BOND, those who complete a TWP. Only 7.1 percent of BOND participants completed a TWP during the first five years of the demonstration, so the program's impacts are meaningfully larger as a share of that directly affected population.

BOND's opposite-signed impacts on the share of beneficiaries working above BYA (positive impact) and on the share working above two and three times BYA (negative impacts) illustrate that a given change to the SSDI the benefit payment formula can provide different a different incentive regarding the choice of whether or not to work from the choice of how much to work conditional on working. To see this, first consider how the addition of the $1 for $2 benefit offset affects the incentive to work above BYA for an individual who in the absence of the offset would not have worked above BYA. For an individual in this circumstance, BOND provides a substantial increase in the incentive to work and earn above BYA by reducing the high implicit tax on the first dollars earned above BYA. For an individual who not would have worked above BYA in the absence of the offset, work and earnings should tend to increase due to a substitution effect. The increase in the share of the Stage 1 treatment group earnings above BYA relative to the control group is consistent with this prediction.

In contrast, for individuals who would have worked even in the absence of BOND, the benefit offset structure provides an incentive to reduce the number of hours worked. Consider an individual who would have worked with earnings in the offset region above BYA in the absence of BOND. For this individual, the introduction of the BOND offset increases monthly benefits paid if work behavior is held fixed, and the $1 for $2 structure increases the marginal tax rate on earnings from 0 percent to 50 percent. The "windfall" of higher benefits with no change to work behavior should tend to reduce labor supply (hours worked) through an income effect. Additionally, the 50 percent marginal tax rate under BOND compared to zero marginal tax under current rules will tend to reduce labor supply through a substitution effect. These responses tend to lower earnings conditional on working above BYA. The decrease in

[30] Somewhat similar findings were obtained in the Benefit Offset Pilot Demonstration (see Weathers and Hemmeter 2011).

the share of the Stage 1 treatment group working above two times BYA and above three times BYA relative to the control group is consistent with these predictions.

The results for BOND Stage 2 were similar in many ways to those of Stage 1. Relative to the control group mean of $49,633, the Stage 2 treatment group received average total benefits over the five-year window that exceeded those of the control group by $1,791 for the treatment group that received standard counseling and by $1,997 for the group that received enhanced counseling, both statistically significant at the 5 percent level. The treatment group's earnings were not statistically significantly different from the control group's earnings.[31] Both Stage 2 treatment groups were more likely than the control group to have worked at least one year with earnings above the BYA. In contrast to Stage 1, the Stage 2 treatment groups were not statistically significantly less likely to earn above 200 percent BYA and 300 percent relative to the control group, suggesting the income effects generated by BOND had smaller consequences for the recruited volunteer with an expected higher interest in working. Finally, the treatment arm that received enhanced work incentives counseling did not have statistically significantly different benefits or earnings from the first treatment arm that received standard counseling.

A somewhat complicating factor in the interpretation of the results is that members of the BOND Stage 1 and Stage 2 treatment groups had until 2017 to complete the TWP and begin a BOND participation period, so those who entered the period near that end date offset could receive offset payments through 2022. As of the end of 2020, the number of participants still generating payments was very small.

BOND is important because it is the first effort involving a nationally representative sample to address the cash cliff problem that has drawn widespread criticism from the research and advocate communities. The results are not encouraging, however, because smoothing out the cliff can both increase and decrease work effort, and the net effect was approximately zero. BOND also had other difficulties. One was that the 12-month accounting period was difficult to implement, with the consequence that it is not clear whether the beneficiaries experienced the benefit offset in the way it was intended to work (Wood and Goetz Engler, Chapter 9 in this volume). In addition, participants in the treatment groups did not understand the benefit offset very well. However, the hypothesis that the relatively modest impacts of BOND were driven by a lack of understanding of the offset rules is not supported by the finding in Stage 2 of similar earnings impacts for the treatment arm that received

[31] However, employment effects for Stage 2 were larger than for Stage 1: 2–3 percentage points on the probability of any employment during 2012–2015 (Stage 2 control group mean was 52.5 percent). Also, the significance of the mean earnings impact was sensitive to whether multiple comparisons adjustments were made to the standard errors. When no such adjustment was made, the mean earnings difference was statistically significant at the 5 percent level for the first treatment arm.

enhanced work incentives counseling and the treatment arm that received standard counseling.[32]

Promoting Opportunity Demonstration (POD)

Similar to BOND in several ways, POD was mandated by Congress in the Bipartisan Budget Act of 2015. POD began in 2018 and will conclude in 2021. POD studies the impact of replacing the SSDI cash cliff with a formula that reduces benefits by $1 for every $2 earned above a particular threshold. The POD sample consists of 10,070 volunteers, ages 20–62 over eight sites, who were randomly assigned to three equally sized groups: two treatment groups and a control group that faces the normal SSDI program rules. The volunteers could be current-pay beneficiaries or individuals whose benefits had been suspended for working over the SGA level. They could be concurrent with SSI but could not have a pending work continuing disability review (CDR). The duration of the treatment is from the time of enrollment to the end of the demonstration in 2021.[33]

The POD intervention differs from that of BOND in several ways. Most importantly, POD eliminates the TWP and EPE.[34] In addition, the benefit offset is operated on a monthly rather than yearly accounting period, which likely simplifies the program from both the operators' and participants' points of view. Combining both of these, from the first month of POD, beneficiaries face a benefit formula that reduces monthly benefits by 50 percent of any monthly earnings beyond a threshold called the TWP threshold, which is lower than the SGA level. If impairment related work expenses (IRWE) exceed the TWP threshold, the offset is applied to earnings above IRWE (with IRWE capped at SGA). For one treatment group (T2), there is an additional rule that a beneficiary loses eligibility for SSDI if earnings for 12 months are high enough to drive benefits to $0. For the other treatment group (T1), eligibility is not contingent on past earnings in this way, and benefits return in any demonstration month where the beneficiary's earnings are sufficiently low (i.e., the same benefits formula is applied every month). Finally, all participants in POD are volunteers, while BOND Stage 1 involved a larger nationally representative sample of beneficiaries that was not restricted to volunteers.

The Mathematica interim evaluation report (Mamun et al. 2021) reports that through 2019, the first year after enrollment was completed, a higher fraction (24 percent) of those in the treatment group are using the offset than was the case for the

---

[32] Difficulty in understanding work incentives rules is a general problem, even in the current-law program.

[33] See the POD design report by Wittenburg et al. (2018) for additional details.

[34] In addition, participants were told that their work and earnings during the demonstration would not count against TWP or EPE accumulations after the demonstration ended, and that after the demonstration ended they would return to the same TWP and EPE status they had prior to enrollment.

Stage 2 (recruited participants) treatment group in BOND (14 percent used the offset in at least one of the first three years). This could be for several reasons, the most likely being the elimination of the TWP. But, among other possible explanations, the higher take-up could also be a result of the quicker responsiveness inherent in a monthly accounting period than in an annual one, and an easier-to-understand benefit formula.

The impact estimates from the interim report do not find evidence that POD had an impact in its first year on any the four primary outcomes of interest (Mamun et al. 2021). During 2019, the treatment groups (pooling T1 and T2) had nearly identical average earnings, probability of substantive employment (earnings above the SGA level), annual SSDI benefits, and total annual income to the control group. The treatment group experienced a substantial increase in employment-related activities such as job seeking relative to the control group, suggesting that impacts on primary outcomes could occur in later years. The report also highlights some implementation challenges. About 6 percent of the treatment group withdrew from the demonstration in the first year, and survey results found that less than half of respondents from the treatment group understood the offset rules.

There are a number of issues with POD that will complicate the interpretation of its findings. One is that some members of the treatment group will be made worse off by POD than they would have been under current-law rules. The clearest case of this is that current rules allow unlimited earnings with no benefit reduction during the TWP, whereas POD reduces those benefits for individuals working over the TWP threshold. In addition, because the TWP threshold is about 75 percent of the SGA threshold, those who would work just below the SGA level under current-law rules will receive a lower benefit under POD, as well. These features interact with the provision that members of the treatment group can always opt to return to current-law rules if they wish. It is possible that those who realize they are worse off under POD than under current law might opt out of the demonstration. This possibility will accurately generalize to a permanent national program that allows people to opt in and opt out in a similar fashion, but not to a program that allows opt in without the option to later opt out.

There is also the overarching question of whether the same opposite-signed effects on work effort that occurred under BOND will also occur under POD. Standard analyses of the different groups involved implies that those opposite-signed effects should indeed occur. Whether they will occur, and at what magnitude, will be revealed by the findings.

One unique aspect of POD is the T1 group, which will be able to work above the TWP threshold for the entire period of the demonstration without jeopardizing eligibility for benefits. Much of the work disincentive in the SSDI program comes from the TWP and EPE structure, which makes termination of benefits an eventually likely outcome for anyone who completes a TWP. That threat is removed for the T1 group during the demonstration, and it will be interesting to see how they respond.

Whether the demonstration period includes a sufficient amount of time to adequately test for this effect is also an open question.

State Partnership Initiative (SPI)

The SPI was a collection of state-level demonstrations fielded from 1998 to 2004 intended to draw SSI and SSDI/SSI concurrent beneficiaries into employment service through Workforce Investment Act One-Stop Career Centers. SSA reviewed state-level programs in 2001 and provided funding during 2001–2004. In all, 12 state-level demonstrations (in California, Illinois, Iowa, Minnesota, New Hampshire, New Mexico, New York, North Carolina, Ohio, Oklahoma, Vermont, Wisconsin) received funding under the SPI umbrella. Though many of the major components of SPI involved employment services, we include SPI in this section of our review covering demonstrations focused on financial incentives because an important component of SPI involved modifying financial incentives related to work. (We discuss it in the next subsection.[35])

The 12 SPI projects all aimed at increasing access and use of employment services for SSI and concurrent SSI/SSDI beneficiaries who demonstrated an interest in work, and states differed in the combinations of employment services they provided. The SPI projects all recruited SSI-only and concurrent beneficiaries through One-Stop Career Centers, which naturally screened for an interest in working. The demonstrations created by participating states provided different bundles of services aimed at removing employment barriers. In varying combinations across demonstrations, these services included benefits/work incentives counseling, case management, placement assistance, job training services and supports from local mental health and developmental disability service providers, workplace accommodations, job service vouchers, psychosocial rehabilitation, peer mentoring, situational assessment, Medicaid waivers, One-Stop Center services, and outreach to employers.

For the most part, SPI was implemented in a manner that did not involve randomly assigned control groups, which limits the lessons that can be drawn. Only 4 of the 12 programs constructed comparison groups, and only 1 of those 4 used random assignment. Uneven data collection for these comparison groups relative to data collection available for program participants provided additional obstacles to formally evaluating these programs. Of the four programs that defined comparison groups, three of the states' treatment groups saw larger employment increases than their comparison groups, and one state's treatment group saw smaller employment increases than its comparison group. For New York, which had an experimental design, Peikes, Moreno, and Orzol (2008) compared the experimental and non-experimental outcomes. They found the two methods gave results that differed in both magnitude and direction,

[35] See the SPI *Conclusions* report by Kregel (2006a) for additional details on the SPI demonstration, and see the final evaluation report by Kregel (2006b) for a focus on this waiver component of the broader SPI demonstration.

likely because unobservable characteristics, such as motivation, locus of control, and health status were more important than the observable characteristics that were available for the treatment and comparison groups.

SPI's SSI Work Incentives Demonstration Project

Though SPI focused primarily on providing employment services, 4 of the 12 SPI state programs participated in an additional demonstration aimed to reduce the financial disincentives to work generated by the normal SSI benefit rules. Also known as the "SSI Waiver Demonstration Project," the SSI Work Incentives Demonstration Project served a recruited subgroup of the participants in the SPI demonstration in the four participating states (CA, NY, VT, WI). The project took place during 2001–2004, beginning later than the main job-services component of SPI. Across the four participating states, programs recruited 1,918 participants. All participants were either SSI recipients or SSI/SSDI concurrent beneficiaries.

The project provided four interventions. First, the waiver provided a "one-for-four" BRR. Normal program rules reduced SSI benefits $1 for every additional $2 earned, after an exemption for the first $65 of monthly earnings. Under the waiver, this 50 percent implicit marginal tax rate on earnings was reduced to 25 percent.

Second, the waiver provided more favorable treatment of unearned income related to work activity, examples of which include Unemployment Insurance, workers' compensation, and state disability benefits. Normal SSI rules exempted $20 of unearned income and applied 100 percent BRR for any additional unearned income. For waiver participants, this form of income was treated as earned income, subject to the higher $65 exemption and subject to the reduced one-for-four BRR beyond the exemption.

Third, the waiver allowed the accumulation of assets in an "independence account." Under this provision, participants were allowed savings of up to 50 percent of gross earnings up to $8,000 per year in a checking/savings account that was excluded from the SSI asset test. The exemption lasted until the end of the waiver program (September 30, 2004) and allowed accumulated assets to be spent down over a 24-month period.

Fourth, the waiver suspended medical CDRs in certain cases. Medical CDRs were suspended for SSI-only waiver participants classified as "medical improvement possible" or "medical improvement not expected."

Unfortunately, the same problem of lack of credible comparison groups that impeded the evaluation of SPI proved even more serious for evaluations of the SSI Work Incentives Demonstration Project. Of the four states, only New York constructed a comparison group (without random assignment), and data sufficient for an evaluation were not collected for the group.

Demonstrations Modifying Vocational Rehabilitation or Other Employment Services

Project NetWork

Fielded between 1992 and 1995, the Project NetWork demonstration studied the effects of intensive case management and employment services on return-to-work outcomes. The demonstration recruited volunteers who were SSDI beneficiaries or SSI recipients or had applied for SSI. It randomly assigned eligibility for services from four separate case management models, with each model implemented at two of eight demonstration sites. Recruited participants were randomly assigned to either a treatment group that was eligible for case management services following the model associated with their demonstration site or to a control group. All participants, including those in the treatment group and the control group, received temporary exemptions to several of SSDI's normal rules regarding the consequences of earning above the SGA level (12 months of work were shielded from the SGA or TWP calculations, and rules were modified regarding CDRs for SSI beneficiaries in Section 1619 status).[36]

Each of the four case management models included in the demonstration provided more intensive services than were typically provided by state VRs. The models differed primarily by the agency or organization that provided the services. Model 1 offered case management from SSA staff. Model 2 offered case management from private rehabilitation organizations contracted by SSA. Model 3 offered case management provided by state VR agency staff stationed in SSA offices. Finally, model 4 offered less intensive referral management services provided by SSA staff. A total of 4,160 participants assigned to a treatment group were eligible for one of the four models.

The main outcomes of interest were earnings and SSDI/SSI benefit receipt and total benefits. Compared to the control condition, the treatments overall increased service receipt, but had only small, short-lasting effects on earnings and no detectable impact on benefits. The treatment groups experienced a modest increase in earnings relative to the control group in years one and two of the demonstration, but their average earnings were not statistically significantly different from the control group by year three.

Ticket to Work

Passed into law in 1999, the Ticket to Work (TTW) program was a reform intended to promote work among SSDI beneficiaries and SSI recipients by introducing a new reimbursement system for employment service providers that was made available both to existing state VR agencies and, marking a change, to private

[36] See the evaluation report by Kornfeld et al. (1999) for additional details on Project NetWork.

providers that the program calls Employment Networks (ENs). Though TTW is a national program, its rollout was staggered in two ways that generated quasi-experimental variation in exposure to the program in its early years that facilitated evaluations of the program's effects.[37]

The idea of TTW is to compensate employment service providers when their services help move participants back to work and off the benefit rolls. Once the TTW program was fully rolled out, all SSDI beneficiaries and SSI recipients receive a "ticket" that they can assign to a participating employment service provider, either a state VR or a private EN, which enrolls the participant in the organization's service plan. The VR or EN can choose to be compensated in one of two ways: If the provider chooses outcome-only reimbursement, the provider begins receiving payments from SSA only if and when the participant stops receiving benefits because of work. If the provider chooses milestone-outcome reimbursement, the provider receives a smaller stream of payments if and when the participant stops receiving benefits because of work, but the provider also receives payments as the participant reaches earnings milestones, representing progress toward program exit, while still receiving benefits. VRs have the option of using the traditional cost reimbursement system that was available prior to TTW, and most tickets are assigned under this option. Under cost reimbursement, the VR can be reimbursed for their service costs if the client/beneficiary reaches SGA employment in 9 months within 12 consecutive calendar months. The VR can choose on a case-by-case basis the cost reimbursement option or their selected payment method under the new options (milestone-only or outcome-only).

Early evaluations of TTW found that payments to service providers under the program's original rules were typically far less than the providers' cost of providing services (Thornton et al. 2007, ch. IX). Addressing this shortfall, a set of reforms in 2008 modified the rules governing reimbursement to service providers in several ways that removed risk and increased expected payments to providers, primarily by front-loading the reimbursement schedule to generate payments earlier in the return-to-work process. First, the reform shortened the payment period for outcome payments for SSDI clients under both milestone-outcome and outcome-only reimbursement so that ENs could potentially receive full payment within 36 months compared to 60 months under the original regulations. Second, the reform added milestones that triggered payments earlier in the return-to-work process under the milestone-outcome option. Third, payments under the milestone-outcome option were set closer to the larger outcome-only payment amount. Fourth, the reform increased payments for services to SSI-only recipients to be closer to the higher amount for SSDI beneficiaries.

The evaluations of TTW have focused on the three main outcomes related to the program's original goals; receipt of services, earnings, and benefits paid (Thornton et

[37] See the final evaluation report by Thornton et al. (2007) for additional details about Ticket to Work.

al. 2007). The quasi-experimental comparisons made in the TTW evaluation examine in the program's early years, because the comparisons are generated by the program's staggered rollout. The rollout was staggered for three groups of states, with Phase 1 states first mailing tickets in February 2002, and Phase 3 states first mailing tickets in November 2003. Mailings were also staggered within groups of states over 10-month windows by last digit of participants' Social Security number (SSN), providing a second source of variation in exposure to the program across individuals that could be studied for program evaluation.

Two main estimation strategies were used to evaluate the early impacts of TTW. The first evaluations conducted by Mathematica used a difference in differences strategy based on a comparison of individuals in the early-rollout Phase 1 states to individuals in the later-rollout Phase 2 and three states. These evaluations found that TTW increased receipt of services relative to the prior VR-only environment but did not find impacts on earnings or benefits paid (Thornton et al. 2007; Wittenberg et al. 2007). This estimation strategy has both advantages and limitations. The main advantage is that the method estimates the intent to treat impact of the policy on the full population of SSDI beneficiaries. The weakness is that the assumption under which the approach is valid, namely that the Phase 1 states would have exhibited similar trends as Phase 2 and three states in the absence of the policy change, is a strong assumption that is not supported by several placebo tests presented in the evaluation. The difference in differences impact estimates were therefore interpreted cautiously.

A second estimation strategy used by Stapleton, Mamun, and Page(2014) exploited a cleaner source of quasi-experimental policy variation, the staggered mailing of tickets to individuals within states following TTW's initial rollout. Comparing individuals who were mailed tickets in earlier versus later months based on the exogenous assignment of mailing dates, Stapleton and colleagues arrive at a similar conclusion that TTW caused an increase in service receipt did not have an effect on the number of months with no benefits paid due to work. The main limitation of this evaluation approach is that the impacts that are identified are very short-term effects for the subgroup of individuals who are "compliers" in the sense that their participation decision was influenced by the month that their ticket was mailed, a population who may be affected by the program differently than the full eligible population. The advantage is that the exogenous variation in mail dates allows for credible estimation of this particular causal effect of the policy. Acknowledging these caveats, neither estimation approach finds any evidence that TTW had detectable impacts on exit for SSDI. The 2008 reforms to the reimbursement structure were not rolled out in a way that generated natural comparison groups. However, during 2007–2010 the number of ENs accepting at least one ticket nearly doubled from 818 to 1,600 (Schimmel et al. 2013), suggesting that entry of service providers responds strongly to financial incentives.

Ohio Direct Referral Demonstration (ODRD)

The ODRD is an effort sponsored by SSA and the state of Ohio to test a more aggressive approach to the use of VR services. In the normal operation of the SSDI and SSI programs, a large fraction of beneficiaries and recipients do not make use of those services, possibly because they are not aware of them. ODRD instead makes a direct referral to VR, with the goal of increasing the rate of usage of VR. The population enrolled in the demonstration are ages 18–19 and either are SSDI beneficiaries and SSI recipients who are undergoing a redetermination of eligibility or are applicants to SSDI or SSI. This study population was selected because young individuals could be more likely to benefit from VR services, because those who are undergoing redetermination are necessarily reevaluating their employment options, and because making use of VR services during the application phase could lead to better employment outcomes whether the applicant receives an award or not.[38]

Enrollment in ODRD began in January 2020 but was paused in March 2020 because of the COVID-19 pandemic. It was restarted remotely in July 2020. The goal was to enroll 750 participants in the demonstration, and for the treatment to be one year in duration, with individuals randomized into either the treatment group (who are referred to VR) or the control group (who are not referred to VR) based on the terminal digit of their SSN. The outcomes of interest are whether the individuals use VR services, whether they remain on SSDI or SSI at a later date, and employment levels.

There are no results yet from ODRD, but it is one of several demonstrations that seek to intervene early, or at least at a critical evaluation point, in an individual's engagement with SSDI or SSI. It will be of interest to see whether early intervention has favorable impacts.

Demonstrations Focused on Employment Services for Those with Mental Impairments

Transitional Employment Training Demonstration (TETD)

The TETD was a randomized control trial (RCT) fielded during 1985–1987 studying the effect of providing transitional employment services to SSI recipients with a diagnosis of intellectual development disorder. The outcomes of interest were employment, earnings, use of disregards, and SSI benefits paid. A set of transitional employment services were provided in 13 different sites around the country. More than 13,000 SSI recipients ages 18–40 were contacted and offered participation in the program. About 5 percent took up the offer. Those who took up the offer were those most interested in improved employment outcomes, and they were disproportionately

[38] See Wittenburg and Livermore (2020) for a review of interventions for youth. We include this demonstration in our review because it concerns changes in the application process for SSDI and the adult SSI program, which many of those reviewed by Wittenberg and Livermore do not. SSA (2020a) provides additional details on the design of ODRD.

young in age. A total of 745 TETD participants enrolled, most with little recent employment history or history of participation in VR. Participants were randomly assigned to a treatment group (*N*=375) who received services or to a control group (*N*=370) who did not.[39]

The treatment group received employment services designed around three goals: (1) placement in potentially permanent competitive jobs, (2) provision of specialized on-the-job training to be phased out over time, and (3) postplacement support and services to support job retention. All sites were required to provide the same basic set of services, but the specific service packages and associated costs varied across sites, with average costs estimated to be $5,600 per person, with a site maximum of $14,000 per person. All services were time limited (transitional). The control group members were provided no services but were free to seek employment-related services on their own.

The control group engaged in few employment activities other than sheltered workshops, but about two-thirds of the treatment group were placed in at least initial jobs. A large fraction of these jobs were stabilized into longer-term employment. Many of the jobs were "community" jobs, which were integrated with the regular workforce, and a majority of participants in these jobs had a coach or training program assistance and so were supported in their employment. The specific nature of the impairment appeared to have little effect on these outcomes.

Quantitatively, the TETD services led to an 85 percent average increase in earnings over a three-year period relative to the control group (whose earnings also increased over the demonstration period), with some evidence of fade out of this impact over time.[40] Over a two-year period, the treatment group showed 10 percent greater use of employment services, a 1 percent increase in supported employment, and a 1.5 percent decrease in the use of sheltered workshops relative to the control group. Additionally, the treatment group experienced an 18 percentage point gain over the control group mean of 56.7 percent as ever having been employed, an increase of 2.5 months in time employed, and a gain of $1,500 in annual earnings (100 percent increase over the control group mean of about $1,500). These employment impacts were accompanied by only a small decline in SSI benefits because some of the increase in earnings experienced by the treatment group was disregarded rather than being subject to the 50 percent BRR.

The outcomes varied widely across the sites. Whether this was because of differences in service delivery, variation in the local labor market, or some other factor could not be determined. However, an examination of the different sites suggested that the impacts were greater in those that emphasized placing participants in potentially

---

[39] See the final evaluation report by Thornton and Decker (1989) for additional details on TETD.

[40] In the first two years, treatment group mean earnings were $1,574 greater than the control group mean of $1,556.

permanent jobs as soon as possible, matching jobs and participants carefully, and being flexible in response to individual participants needs.

TETD is one of the few demonstrations that has targeted SSI rather than SSDI. Its impacts were largely positive, which establishes the potential for employment increases to be possible for SSI recipients. As with many RCTs, the participants were volunteers who were most interested in improving their employment. The large variation across sites, common to many RCTs, is a cause for concern. One issue with TETD is that its findings are 35 years old; whether they would hold with today's caseload is unknown. In addition, new models of providing services to those with mental impairments have been developed since TETD, such as the Individual Placement and Support model discussed next.

Mental Health Treatment Study (MHTS)

The MHTS was a demonstration conducted from 2006 to 2010 to test the impact of a new set of employment services to those on SSDI who had a primary diagnosis of schizophrenia or affective disorder. A central part of the MHTS treatment used an approach to employment services called the Individual Placement and Support (IPS) model, which couples obtaining jobs for participants with disabilities with a suite of associated services. These include employment services, mental health services, benefits counseling, and individualized job supports (participants receive individualized supports as long as they want it in their jobs, and their employers are offered support, as well). The IPS model also aims for participants to obtain jobs in the competitive labor market, to be contrasted with sheltered workshops or supported employment jobs. It seeks to move participants into jobs as quickly as possible after IPS enrollment and without lengthy training. IPS also aims to allow the participants to help direct the course of activities according to their own preferences, rather than strictly following a plan devised solely by the program staff (Frey et al. 2011).

The MHTS demonstration invited a random sample of SSDI beneficiaries ages 18–44 in 23 different study sites. The take-up rate was about 14 percent, with a resulting sample of slightly more than 2,000 enrollees. Enrollees were interpreted as those most interested in employment. They were approximately equally randomized into a treatment group and a control group.

The treatment group received those services included in the IPS model, but also a large additional set of services. These included systematic medication management, integrated behavioral health and employment services, comprehensive health insurance coverage, and nurse coordinator counseling. The medical CDRs were also suspended for three years from the date of enrollment. The treatment group received services for a 24-month period.

The control group received only a manual that listed local and federal resources for persons with mental illness, and their medical CDRs were not suspended. However, they were free to seek any services they might be interested in obtaining.

The impact estimates showed a statistically significant increase in employment, with a 61 percent rate for the treatment group and a 40 percent rate for the control group (Frey et al. 2011). The treatment group had higher earnings than the control group did, although still below the SGA level. Service occupations and sales and office occupations were the most common types of jobs. Predictors of increases in employment were past employment history, health status, and the local unemployment rate. Because earnings were typically below the SGA level, SSDI benefits were not reduced. The treatment also improved participants' mental health.

The MHTS appears to be a successful model for increasing employment among those with certain types of mental impairments. Some regard the IPS model as a successful model and superior to older-style VR (Bond 1998; Bond, Drake, and Becker 2008), although the evidence from MHTS is not definitive on this because the treatment included more than IPS and was limited to certain impairments. The evidence from MHTS is that the major impact is on employment with earnings below the SGA level, rather than higher earnings, and so exit from SSDI should not be expected from this intervention. It also appears that the MHTS treatment, including not only the IPS services but also the rather large number of additional supports and financial subsidies for health insurance costs, is quite expensive—approximately $7,000 per participant. This could make it infeasible financially if the full SSDI population participated, even only those with these types of impairments.

Supported Employment Demonstration (SED)

The SED is an ongoing RCT designed to study the effects of IPS and other services on employment, SSDI and SSI receipt, and mental health of individuals who had applied for SSDI claiming mental impairment but had been denied. The demonstration recruited 3,000 denied applicants ages 18-59 who reported wanting to work or were already working at the time of recruitment. The participants were randomly assigned to one of three groups: two treatment groups and one control group. The control group maintained access to the usual available services and were given an informational handbook. The two treatment groups were provided additional services. The three groups were roughly allocated one-third of the sample each (SSA 2020a).

The first treatment group is receiving what are deemed "basic" services based on the IPS model, as described for MHTS above, but supplemented with behavioral health and employment-related expenses assistance. Care management is also provided to help in coordination.

The second treatment group receives all basic services, plus additional medical services in the form of a nurse care coordinator, systematic medication management, and cost-sharing for medications.

SED recruitment began in November 2018 and will provide services to the treatment group 36 months. The main outcomes of interest include employment, SSDI receipt, SSI receipt, mental health, and quality of life. We expect that the SED final evaluation will include a cost-benefit analysis, so the cost of services is another

important outcome. If successful with job placement, SED could also affect program entry by reducing the probability of a participant appealing the initial SSDI denial or re-applying for benefits at a later date.

SED is of particular interest because it targets those with impairments who have been denied an award. This is presumably a population with less severe disabilities than those who obtain an award, and we know less about the impact of employment services on that population. It constitutes an early intervention, which not only has the chance to assist those who are not yet receiving SSDI but still have major impairments, but it also could delay an appeal to reverse the initial denial (although appeals have important time limits). However, it is also likely to be more challenging to provide services to the study population than those on SSDI already because the SED population is less integrated into the community that provides services to people with disabilities. Connecting them to service agencies with which they might not be familiar, and therefore, for the first time, is likely to be an issue in achieving positive outcomes.

Demonstrations Focused on Expanded Health Insurance Provision

Accelerated Benefits (AB)

The SSDI program awards Medicare coverage to individuals, but only 24 months after they first become entitled to benefits. Some beneficiaries have health insurance coverage from their previous employer, through spousal coverage or other coverage from the family, from Medicaid if they have low income, or other sources; however, about one-fifth have no coverage at all (Michalopoulos et al. 2011). Yet they often have high medical need and make frequent doctor and hospital visits. The AB demonstration was primarily aimed at studying the impact on health status of health insurance access in the 24-month waiting period (and after it). But employment and return-to-work outcomes were also of interest because any improvements in health resulting from insurance coverage could lead to increases in labor market engagement (Michalopoulos et al. 2011).

In AB, about 2,000 SSDI beneficiaries were enrolled between October 2007 and January 2009 in an experiment that randomly assigned subsidized health insurance to approved SSDI beneficiaries before the end of the usual 24 months. Enrollment took place from 2007 to 2009 and the treatment was for 24 months. Participants in AB were SSDI-only beneficiaries (SSI recipients were excluded because they are covered by Medicaid), ages 18–54, who did not have health insurance coverage at the time of enrollment and who had no less than 18 months remaining in their 24-month waiting period for Medicare. The participants were recruited from the 53 largest Standard Metropolitan Statistical Areas.

Participants were randomly assigned to one of two treatment groups or to a control group. The first treatment group (E1) received health insurance coverage for 24 months, spanning the Medicare waiting period. The plan covered hospital, medical,

and drug claims; use of skilled nursing facilities; home health care; prosthetics, vision, hearing, and dental care; and some out-of-network services. Other than a $12 copayment required for the majority of services provided by the plan, participants were not responsible for any premiums or other costs. However, more expensive care such as emergency room services and inpatient care required higher copayments, and limits were placed on some services such as inpatient care for mental disorders, chemical abuse treatment, and use of skilled nursing facilities. The maximum health care benefit available to a participant during the demonstration was $100,000. At the end of the 24-month period, all AB participants transitioned to Medicare.

The second treatment group (E2) received the same health insurance benefit as E1, plus medical care management services; the Progressive Goal Attainment Program, or PGAP® (a self-paced rehabilitation program to improve functioning); and employment counseling services. About 74 percent of the group took advantage of at least one of the three services. The AB intervention did not make any changes to SSDI's normal rules regarding the impact of work on eligibility, such as the TWP, EPE, and work CDRs.

Members of the control group received the usual SSDI program benefits and services including its health insurance component. As in the usual program as described above, any health insurance coverage would come from sources outside of SSDI, including Medicaid for beneficiaries who meet the Medicaid means test, coverage from a beneficiary's previous employer, spousal coverage, or Affordable Care Act coverage from an exchange plan.

There are several publications reporting the impact of the AB intervention (Michalopoulos et al. 2011; Weathers et al. 2010; Weathers and Stegman 2012; Bailey and Weathers 2014; Weathers and Bailey 2014). The main impact estimates showed that AB increased health care use and reduced unmet medical needs relative to the control group (over the course of the demonstration, health insurance coverage increased for the control group, as well). The treatment groups experienced improved health status relative to the control group. There were no measurable short-term effects on mortality. AB also increased participation in TTW, which might be expected to increase employment.

For employment, there were no statistically significant differences between E1 and the control group. There were no statistically significant differences between E1 and E2 in employment in the first year in AB, but there was an increase in employment in the second year for E2 relative to E1. There were no statistically significant differences across groups by the third year, by which point participants from all groups had typically reached Medicare eligibility.

These impacts of E2 relative to E1 imply that the provision of the medical services, rehabilitation services, and employment counseling services had an impact even though health insurance coverage and improvement in health per se did not. The timing of the results suggests that members of the E2 treatment group used the services in the first year, then in the second year benefitted from those services by increased

employment. However, the lack of effects in the third year suggests that there was no long-term impact of the increased employment in the second year on later employment levels.

The AB experiment was expensive. During the first year of AB, the cost per participant for E1 was about $31,000 and for E2 about $34,000.

Strictly regarded from a labor market perspective, the AB demonstration shows the importance of the combination of health insurance coverage and medical and employment services to labor force activity of SSDI beneficiaries. However, the lack of a long-term impact on employment suggests that the provision of that insurance and those services, though improving health and other outcomes, is unlikely to have a major impact on employment while receiving SSDI or on exit rates from SSDI. But AB is among those demonstrations that seek to intervene early in a beneficiary's time on SSDI. It also has the more general implication that early intervention in general could be a path worth pursing in future demonstrations (Hollenbeck 2021).

A question about the relevance of the AB demonstration is whether the Affordable Care Act might make its effects different today. The Act provides subsidies for individuals who work and earn enough to put their incomes into the subsidy range. And in states where Medicaid expansion took place, childless individuals are more likely to be covered. This might affect the impact of a health-insurance subsidy program like AB.

LESSONS LEARNED

Our review of the 11 demonstrations in the last section leads us to draw several conclusions. These conclusions will lead us to make a number of recommendations for new demonstrations, which we enumerate and discuss in the next section.

1. Most of the efforts to increase employment, earnings, and labor force engagement of SSDI beneficiaries have been disappointing.

The effort to remove the cash cliff tested in the BOND experiment led to higher employment among Stage 2 participants of 2 to 3 percentage points and only a very small increase for Stage 1 participants (0.36 percentage points). BOND also had a null effect on average earnings of Stage 1 and Stage 2 participants. For the Stage 1 group, earnings increased just above the SGA level and decreased among those at higher earnings. For the Stage 2 treatment group, earnings increased just above the SGA level but experienced only small increases at higher levels, leading to a small overall earnings gain.

These small earnings effects could be the result of opposite-signed incentives at different levels of earnings but also possibly the result of problems in implementation and of a treatment that too many participants did not understand. Efforts to add counseling to participants to help them understand the benefit offset had no effect on

earnings but did increase understanding[41] The Project NetWork demonstration had only small increases in employment and earnings. The TTW evaluation showed increases in service use but only a small fraction of participants took their tickets to a provider, suggesting that little subsequent impact on employment is likely.

Many of these demonstrations faced difficulties in implementation, and in several cases participants appeared not to understand the work incentive rules well, as we mention below. Technically speaking, we cannot rule out that those problems led to the lack of major employment impacts. However, on the key underlying question of whether the low employment rates of SSDI beneficiaries are the result of low residual work capacity, or the work disincentives in the SSDI rules, we can safely say that we have no evidence thus far that the disincentives are the major problem. Analogously, we can say that these demonstrations do not provide evidence of the existence of a large residual work capacity that is untapped and only not exercised because of work disincentives.

2. The demonstrations for SSDI beneficiaries with mental impairments were an exception and showed more favorable effects.

TETD showed favorable impacts on employment and earnings, as did MHTS. MHTS took place much later in time than TETD and showed that applying the IPS model combined with a large number of additional services could raise employment and earnings. It is important to note that MHTS did not provide evidence on the impact of IPS alone, only on the impact when it is supplemented with substantial additional services. It should also be noted that though MHTS did improve employment, the earnings gains were not sufficient to reach the SGA level, so we should not expect this reform to increase exit from the program.

Another issue with MHTS is that the large number of services provided makes it quite expensive, and it is consequently questionable whether it could be applied to the SSDI caseload at a large scale. Still the concept of trying to get beneficiaries to engage in a large number of services simultaneously because the impacts of the services could complement and reinforce one another—that is, that the impact of the total could be greater than the sum of the impacts of the subcomponents if each had been implemented separately—might apply to other interventions.

3. Where there are favorable SSDI earnings impacts, earnings rarely rise to the SGA level.

Most increases in earnings do not rise to the SGA level, in demonstrations both for beneficiaries with non-mental impairments and for those with mental impairments. This necessarily means that increases are small. Whether this is simply because beneficiaries have too many barriers to work or because they intentionally do not want to work at the SGA level because of the risk that it might lead to termination of benefits

[41] See Chapter 8 in this volume for a review of the effect of counseling in SSA demonstrations.

cannot be ascertained. But, again, this general result implies that few additional exits from the SSDI program can be expected from the demonstrations tested thus far. A question for policymakers is how much value they wish to put on increasing employment and labor force attachment even if it does not reduce the caseload and does not reduce SSDI expenditures (indeed, in some cases, it could increase them). Some demonstrations (e.g., MHTS), though not increasing earnings to the SGA level, had favorable impacts on mental health, possibly a result of beneficial effects of working more, so there are other possible outcomes for some of the reforms. Answering this question is outside the scope of our review but is a general question about the goals of reform.

One suggestion that has been made to estimate the maximum work capacity of the SSDI caseload is what is called the "Ultimate Demonstration," (see Gubits et al. 2019), which would offer recipients the opportunity to receive a lower benefit level in return for no earnings restrictions at all and for an indefinite period of time (but medical CDRs would still be conducted).[42] Beneficiaries would have the option of returning to the standard program if they wished. One issue with the Ultimate Demonstration is that the chance that it would receive congressional approval is probably low, so the expenditure of SSDI trust fund dollars on a reform that is unlikely to be accepted might not be desirable. But a more important problem is that such a reform would, like all reforms that essentially lower the BRR, have ambiguous effects on average work effort because some beneficiaries would be induced to work less, not more, because of the income effects of the reform (namely those who would work above the SGA level in the EPE). We know this to be a possibility because it seems to have occurred in BOND.[43]

4. There are essentially never increases in exits from these demonstrations and rarely reductions in SSDI expenditures.

We have already noted this as a corollary of the results showing that earnings rarely rise to the SGA level. In fact, many of the demonstrations increase benefits paid, not decrease them. Again, as just noted, the importance of this conclusion depends on whether the goal of these reforms is to lead beneficiaries to leave SSDI and become self-sufficient, or just to encourage them to engage in work while on SSDI, or what weight each goal is given. Nevertheless, even if the majority of beneficiaries who respond to the demonstrations' reforms do not leave SSDI or reduce expenditures, it is disappointing that even a small fraction do not leave.

---

[42] Another option is not to reduce the benefit at all when offering the program, which could yield a quite different result.

[43] Even if not likely to be politically feasible, the Ultimate Demonstration would provide information on the extent of residual work capacity of the beneficiary population.

5. SSDI financial incentives do not work so well.

Despite the general objection by many to the SSDI cash cliff, smoothing it out does not appear to increase earnings and could decrease earnings because there are offsetting impacts to the increase in earnings of beneficiaries initially below the cliff. Smoothing the cliff could be desirable for other reasons (such as a more equitable treatment of those just below and those just above the SGA level), but increasing average work effort should not be one of them. Having said this, we must repeat the caveat that we noted in Lesson 1 above that difficulties in implementing and lack of understanding of the rules could have contributed to the lack of effects of smoothing out the cliff. But in this case, there are theoretical reasons, as we have emphasized, that net effects can be small or wrong-signed. Yet another possibility is that even after smoothing out the cliff, the BRR is still too high. We discuss this further in our recommendations for new demonstrations.

6. Implementation and operational constraints are real.

Some of these programs (BOND is the clearest example) have treatments that require SSA to make changes in the program that are difficult to implement and would constitute a barrier if they were to be scaled up (Chapter 9). This is not too surprising given the evidence that SSA has difficulty implementing the work incentives rules under current law (Wittenburg et al. 2012, 7) and the reforms in the demonstrations are often even harder to implement.[44] This suggests that the design of future interventions should give simplicity of implementation high priority.[45] At the same time, it should be recognized that the underlying problem is that there is a real tradeoff, because trying to provide the optimal incentives to SSDI beneficiaries and SSI recipients can necessarily lead to complexity. Something might need to be given up to achieve simplicity; it is not costless.

7. Understanding the treatment is also key.

Demonstrations that have complex treatments not only are difficult to implement but also are difficult for participants to understand. Treatments that are not understood should not be expected to have much impact.[46] If implementation difficulties occur at the same time, participants are likely to be even more confused about the rules. Again, there is a tradeoff, because designing rules to have optimal work incentives can lead to complexity. So again, something might have to be given up.

As we have noted previously, both implementation difficulties and lack of understanding also hinder the interpretation of the results on program impact, because

---

[44] For other examples of implementation difficulties, see GAO (2017) and SSA (2016).

[45] POD attempted to do that to some extent.

[46] Sometimes the rules create uncertainty, as in the cases of BOND and POD where members of the treatment group may have had concerns about overpayments and having to return money later.

it cannot be known whether the impact is a result of the treatment itself, implementation deficiencies, or lack of understanding. Nevertheless, the question is whether the rules that have been tested in existing demonstrations could be implemented better or if they could be explained to beneficiaries and recipients better than they have been. Though there is no rigorous demonstration evidence on this—because we do not have demonstrations that just vary the effort put into implementation or the effort put into helping participants understand the rules while holding the actual rules fixed[47]—we suggest that it could be that the rules themselves are simply too difficult to implement and understand.

We should also note that there is also a question of whether beneficiaries understand the current-law incentives of the SSDI program (Wittenburg et al. 2012). What this implies is that a demonstration needs to have rules that are clearer to understand than current-law rules, comparatively speaking. Having easier-to-understand rules could be part of the treatment in addition to the other reforms, and the experimental-control difference in outcomes would reflect both.

8. Only a small number of SSDI beneficiaries take up most of the programs.

The take-up rate of the offer of services in the treatment group, as well as the take-up rate of the offer to participate in demonstrations in the first place, is typically quite small. This could be because the treatments being tested are not attractive enough to participants for whatever reason, but it also could reflect that many fewer beneficiaries have substantial residual work capacity than has been thought previously. We suggest that expectations for the fraction of SSDI beneficiaries who have sufficient residual work capacity to take advantage of the programs should be lowered, although recognizing that the size of residual work capacity in the SSDI caseload is still uncertain. Maestas, Mullen, and Strand (2013), for example, have suggested that for those SSDI beneficiaries who had been just on the margin of being accepted into the program (about 12 percent of beneficiaries), residual work capacity is quite high. Of this marginal group, those who were not awarded benefits were 28 percentage points more likely to be employed two years post-determination than otherwise identical applicants who were awarded benefits.[48] However, whether the other 88 percent of beneficiaries have anything close to that level of capacity is questionable.[49]

---

[47] An exception is BOND Stage 2, which included two treatment arms that differed only in the intensity of work incentives counseling. Those receiving enhanced work incentives counseling had similar employment outcomes to those receiving the standard version, consistent with the idea that the rules themselves likely present a larger barrier than does a lack of counseling effort.

[48] Awards and rejections for the set of applicants involved in this comparison can be thought of as randomly assigned under the study's identifying assumption that a particular component of the assignment of SSDI applicants to disability examiners is as good as random (Maestas, Mullen, and Strand 2013).

[49] The 28 percent reduction is relative to a small base. The results are sensitive to what earnings range is being considered.

9. Many fewer demonstrations have been tested on the SSI population, but the simple evidence from SSI is not promising.

The SSI program should be of greater interest than it has been because it has a 2-for-1 benefit offset already, it has no SGA rules that create cash cliffs, and it does not terminate recipients from the program after sufficient periods of work over the SGA level, as the SSDI program does.[50] But despite these much larger incentives to work, only 7 percent of SSI recipients ages 18–64 did so in 2019 (SSA 2020h, Table 42). Almost all SSI recipients who do, work below the SGA level even though they could continue to work above it without loss of eligibility. This suggests again that financial incentives and the SSDI program structure might be less important barriers to work than just the small numbers of recipients with substantial residual work capacity. A qualification to this conclusion is that the SSI program caseload is somewhat different from the SSDI caseload, and because SSI recipients have less work history than SSDI beneficiaries, SSI recipients could have lower residual work capacity to begin with.[51]

10. Early intervention might have promise.

We reviewed three demonstrations that intervened early in the SSDI application and post-award time period (AB, ODRD, and SED).[52] Two are ongoing, but the one that has been completed (AB) had favorable employment and earnings impacts from early health insurance provision combined with rehabilitation and employment services during the first 24-month period of SSDI eligibility. Though the effects were not long-lasting, this could be simply because the treatment ended at 24 months, when both treatment and control group members began to have the same health insurance coverage. This suggests that other early interventions might be worth considering for future demonstrations, although clearly AB is only one demonstration, and there have been many more early interventions proposed (for a list of several, see Chapter 5 in this volume).

SUGGESTIONS FOR FUTURE PROGRAM REFORMS AND DEMONSTRATIONS

Here we list a few ideas for future program reforms, based on those lessons learned combined with our general sense of what the barriers are to successful interventions. We divide our ideas into two groups, the first consisting of ideas for

---

[50] Note that both the YTD and SPI did test reforms on SSI, however.

[51] The information in *DI & SSI Program Participants: Characteristics & Employment, 2015* (SSA 2020d), especially the section "Work Activity Before and After Award" (pp. 5–11), is useful on the work aptitude of SSI recipients (and SSDI beneficiaries).

[52] Pre-application interventions include Retaining Employment and Talent after Injury/Illness Network (RETAIN), Promoting Work through Early Interventions Project (PWEIP), and one Youth Transition Demonstration (YTD) site, described in other chapters in this volume.

substantive programmatic changes and the second consisting of ideas for addressing design issues in demonstrations in general.

Programmatic Changes

Earned Income Tax Credit. There is a vast research literature on the effect of financial incentives on the work incentives of transfer program recipients of various kinds, with the most common topic being the effect of lowering the BRR imposed on increases in recipient earnings. The overall conclusion of that literature, in our view, is that the effects of such a lowering are small if not negative. The results of the BOND experiment are roughly consistent with that literature. Aside from the issue of opposite-signed effects, many analysts also believe that a 50 percent tax rate on earnings is far too high. Such tax rates are rarely present in most countries' income tax programs, for example, and tax rates that high are mostly considered onerous. Transfer program recipients might be no different from the rest of the population in this regard, and in fact, they might be more conscious of benefit losses of that magnitude if their marginal utility of income is greater than that of higher-income individuals.

The most successful financial reform in other transfer programs in the last few decades has been the Earned Income Tax Credit (EITC). The EITC not only does not tax recipient earnings, it subsidizes them, at least when recipients move from nonwork to work and when increasing earnings at low levels. The research evidence has shown favorable impacts of the EITC on individuals who are initially concentrated at nonwork and whose work decisions are sensitive to financial considerations.[53] These favorable impacts are much greater than the effects of lowering the benefit offset rate in a transfer program, presumably because recipients do not like to have their benefits reduced much at all, but they like being rewarded rather than penalized for working.

The EITC necessarily has a phaseout region, and this could easily have disincentives, as any benefit withdrawal does. However, the evidence showing favorable net impacts on earnings from the EITC has been interpreted as implying that the work disincentives of withdrawal of benefits are not as large as the work incentives of the subsidies, possibly because intensive margin elasticities are smaller than participation elasticities. Some researchers have speculated that this is because the subsidy rate in the phase-in region is high (up to 40 percent of earnings) whereas the BRR in the phaseout region is low (about 20 percent). This suggests that some consideration could be given in a future demonstration to providing high BRRs in SSDI or SSI at low earnings levels and low BRRs thereafter. The question would be whether SSDI beneficiaries, or some fraction of them, have sufficiently high participation elasticities to respond to such a reform.

Gokhale (2013, 2015) has proposed something similar for SSDI, which he calls the generalized benefit offset. However, his plan is quite complex and differs markedly

[53] In other words, they have high participation elasticities. Single mothers have been identified as one such group. See Hotz and Scholz (2003) and Nichols and Rothstein (2016).

from the EITC. He proposes that workers who earn less than the SGA level have their benefits taxed, leaving them worse off than under current law and discouraging low levels of work. But benefits are extended beyond the SGA level at an increasing rate per dollar of earnings until about three times the SGA level, at which point they are taxed away and end at about four times the SGA level. The benefit schedule is nonlinear, with phase-in and phaseout rates changing for every extra dollar of earnings, which would be difficult for beneficiaries to understand. Gokhale's phase-in and phaseout rates are also symmetric—both about 36 percent—unlike the EITC, which has a very steep phase-in rate and a very slow phaseout rate.

Many details would have to be worked out in the implementation of an EITC in SSDI. For example, basic questions would have to be answered, such as whether EITC would be subject to the same fixed-length period of the TWP (number of months it takes to complete) followed by an EPE. An alternative would be to restructure the TWP-EPE sequence in some way, or to eliminate them altogether as in POD. Presumably, medical CDRs would continue, but whether work CDRs would continue depends on whether benefit termination policies would continue in the same way they are in the current-law program. Another possibility is to combine the basic EITC with other front-door or back-door policies, as described below.

Reducing the prospect of termination. One of the most likely disincentives to work for SSDI is not the cash cliff but the prospect of benefit termination if the beneficiary works too many months earning at the SGA level. Though this requirement has always been central to the structure of the SSDI program, POD will provide some evidence on the importance of this factor. Whereas one treatment group in that demonstration experiences termination from the program after 12 months of earnings, leading to zero benefits, a second treatment group experiences no termination from the program at any future date (during the demonstration) if working 12 or more such months. If the results show larger work levels in the second treatment group than in the first in those 12 months, it suggests that the prospect of benefit termination is playing a large role in the current SSDI program. However, the impact of the policy is likely to be sensitive to how high earnings have to be to reach the zero-benefits point. If that point is too high, for example, too few beneficiaries will ever reach it, and so this aspect of the reform will not be effective.[54]

Any policy of this type that alters the point at which termination takes place is often called a "back-door" policy. Aside from altering the number of periods of earnings that result in termination or the level of earnings that count toward termination, an EITC inducement to leave SSDI might be worth considering. Providing an EITC to supplement the earnings of beneficiaries who leave SSDI entirely could provide an inducement for the same reason that the general EITC described above does—namely, instead of just reducing the penalty of leaving the

[54] The limited duration of the demonstration may also hinder the effectiveness of the difference in termination rules.

program, it can make the beneficiary financially better off by leaving, relative to current law. Such an EITC would presumably have to be time limited.

Early intervention reforms. It is widely believed by labor economists that long periods out of the labor force result in degradation and deterioration of labor market skills and reductions in desire to work and knowledge of working, leading to greater difficulties to return to work later. This suggests that early interventions ("front-door" policies) could have greater chances of success. One example comes from AB, which changed the beneficiary experience in the first 24 months after establishing eligibility. The results from ODRD might yield more information on this approach. SED is rather different, because it targets applicants who have been initially denied eligibility; it is nevertheless related, because it seeks to offer new programs to those initially denied immediately after the decision to deny eligibility has been made.

More generally, the long process that many applicants experience when waiting for decisions and complying with application requirements suggests that more interventions during the application process might be considered. This could involve offering financial incentives to work to those who have not completed an application; or it could involve offering employment services, including counseling, to applicants during the process.[55] Though we have not reviewed demonstrations that test such programs, their appeal is that they keep individuals who have experienced a disabling event connected to the world of work and employment during the application process, which might consequently enable them to re-enter into employment more quickly later.

Hollenbeck (Chapter 5) reviews a number of studies proposing reforms at the application stage (including one that offers an EITC to applicants). Many of these proposals offer assistance to applicants who will be denied eventually by regular SSDI procedures, representing an increased value of applying that might draw in other applicants. The proposals also might increase total SSDI expenditures because a new group is being served. However, if applicants who are eventually denied benefits end up having improved outcomes from the services they receive during the application process, that should be regarded as a social benefit. Further, despite this effect, the goal of such programs would be to start offering employment services to applicants who will eventually receive an award. As just noted, the original idea of early interventions is to not wait until much later, in some period after a SSDI award is made, before beginning to reacquaint beneficiaries with the world of work and to establish and maintain work skills and habits.[56]

---

[55] SED gets at this. Also see "Communicating Employment Supports to Denied Disability Insurance Applicants" (https://oes.gsa.gov/projects/di-denial/).

[56] There have also been many reforms suggested for intervening prior to SSDI application, with the RETAIN project being one of these (Hollenbeck in this volume). We do not cover pre-application demonstrations or reforms in this chapter, especially those testing programs on groups that could differ substantially from SSDI applicants.

Simplification and reduction of uncertainty. SSDI rules are fairly complex even in current law, and not all beneficiaries understand them. Moreover, beneficiaries are likely very uncertain as to how the rules will be applied and what the consequences will be for specific employment and earnings decisions they are considering making. The complexity of the work rules is compounded by a cumbersome process by which SSA applies the rules, usually involving an administrative procedure taken only some time after the applicant has worked and after earnings amounts are documented by an often laborious process (Wittenburg et al. 2012). Demonstrations that greatly simplify the rules and reduce uncertainty of the application of those rules are worth considering. Such demonstrations could succeed only if SSA was able, in fact, to make its application of the rules more certain.

An issue with any demonstration on work-rule simplification is the practical problem that SSA has demonstration authority only to test new rules on volunteers, and the question is who would volunteer for a demonstration that was only about rule simplification and nothing else. Further, while using volunteers is acceptable for demonstrations that would never be applied to the full caseload were they made permanent, work-rule simplification would be imposed on the entire caseload (although only affecting those interested in working). In all likelihood, obtaining volunteers for a work-rule simplification reform might be possible only if it were coupled with some more substantive reform in those rules that volunteers would find attractive to try out. Although this would mean that the effects of simplification versus the more substantive reform could not be separated, that would be acceptable if such a condition is also what a permanent reform would look like.

Smoothing the cash cliff. Further demonstrations to test smoothing the cash cliff should be given low priority. The cash cliff could be smoothed to avoid an inequitable bimodal distribution of benefits above and below the cliff, rather than aimed at increasing average employment and earnings, as mentioned above. However, gradual phaseouts of benefits, whether the zero-benefit limit is at the SGA level or something higher, could easily be accomplished as part of other demonstrations that alter the benefit structure in SSDI, without making phaseouts the focus of the demonstration. Gradual phaseout, for example, would be preferable in a demonstration on an EITC.

Time-limited benefits. Some transfer programs offer applicants a time-limited benefit in return for withdrawing their application (e.g., state Temporary Assistance for Needy Families programs have that option, where it is called a diversion policy). Such programs have been suggested for SSDI (Stapleton, Ben-Shalom, and Mann 2019). For example, offering a substantial cash benefit plus VR or employment services for a fixed term might be attractive to applicants who believe they are very likely to return to work. At the end of the fixed term, participants would have the option of again applying for SSDI. Like the early intervention programs discussed earlier, this reform would have the advantage of starting individuals who would eventually become SSDI beneficiaries on the road to labor force reengagement at an early stage. However, part of the intent would be to lead some of those who might

have otherwise gone onto SSDI to instead recover sufficiently that instead of going onto SSDI at the end of their fixed term, they would be capable of attaining self-sufficiency off SSDI.

We noted above the possible attractiveness of early intervention reform, but a time-limited benefit could have a larger application-inducing effect than treatments that only offer employment support services or other non-cash opportunities. Depending on the size of the benefit, the induced application effect could also be quite a bit more expensive. Finally, operating a benefits program parallel to regular SSDI but for applicants is likely to pose more administrative difficulties than just an employment services program. These issues would need to be weighed before considering a time-limited benefit program for SSDI applicants.

Partial SSDI. SSDI systems in many countries include some form of partial SSDI provided to individuals whose earning capacity is reduced due to a disability but who still have the capacity to earn at a level above the SGA level. Recent proposals for reforms to the SSDI program in the United States have included calls to introduce a partial SSDI component to SSDI (Maestas 2019). Under this type of program, an individual with a disability whose earning capacity drops (but is above the SGA level) would receive a partial SSDI benefit (proportional to the reduction in their earning capacity) but would be allowed to work while receiving benefits. A partial SSDI program likely would need to be accompanied by more comprehensive initial medical assessments to estimate each applicant's earning capacity as a specific fraction of their pre-disability earning capacity, as opposed to the current binary classification of disabled versus not disabled (Maestas 2019).

A role for partial SSDI is likely more appropriate now than when SSDI began due to recent trends in the composition of the SSDI applicant pool. Larger shares of SSDI applications now come from individuals with musculoskeletal and mental impairments, more of whom are likely to have a residual capacity to earn above the SGA level than other applicants (SSA 2020b, Table 21). For individuals with residual earning capacity above the SGA level, partial SSDI provides insurance that is tailored appropriately to the size of their loss while providing incentives more aligned with long-term attachment to the labor force compared to current SSDI. Unlike the current SSDI program, partial SSDI would not make an applicant stop working entirely to demonstrate eligibility, and the applicant would face a smaller financial disincentive to work after being awarded.

As with introducing time-limited benefits, expanding SSDI to include partial SSDI for a population that is currently not strictly eligible for SSDI could have application-inducing effects. Partial SSDI also presents a different set of challenges for the design of an informative demonstration. Careful consideration would be required to determine how to structure a version of partial SSDI that is generous enough relative to the current SSDI program to attract demonstration volunteers while still providing evidence that would generalize to the kinds of partial SSDI programs that Congress might consider enacting. One attractive feature of partial SSDI is that it

does not require individuals to fully withdraw from the labor force in order to apply. For this reason, a demonstration might want to recruit volunteers and implement random assignment as early in the application process as possible, even if return-to-work outcomes of those eventually awarded partial SSDI are important outcomes of interest. Another option would be to draw volunteers from other demonstrations that target the application phase or pre-application phase, such as RETAIN. Alternatively, a demonstration could target current beneficiaries who have entered a TWP or who show an improvement to work capacity on a medical CDR, offering re-assessment for partial SSDI as an alternative to the current program rules.

Demonstration Design Issues

Multiple treatment groups. A single treatment group that receives an intervention with multiple components does not permit attribution of impacts (favorable or unfavorable) to any single component. Some of the demonstrations we reviewed involved evaluations that have two treatment groups, but never more than that. The main issue is the cost of the demonstration if each group is sufficiently large to obtain significant statistical power. One specific case where sufficient power could be retained is where the treatment is scalable; for example, testing various levels of BRRs, various earnings levels at which benefit offset are set to begin, or various accounting periods. By pooling treatment groups with varying levels of the scalable variables and imposing some functional forms on the shape of the response function (for example, by assuming their effect is linear and proportional, as would be the case in linear regression analysis), it is possible for power to be retained in the estimation of that function.

Scalable treatments that can save on statistical power in this way are an example of a type of factorial design, as defined and discussed in Chapters 2 and 3. Also possible is the more general case of factorial designs, where only some combinations of multiple treatments are tested and then the results are used to extrapolate to combinations not tested. We suggest that these and other multiple treatment demonstrations be considered.

Working somewhat against this idea, however, are the results of the MHTS demonstration, which suggest that positive effects may be more likely if multiple treatments are imposed simultaneously. As noted above, SSDI beneficiaries often have multiple barriers to work, and interventions which several components intended to address them all at one time may be more effective than offering each incrementally. Achieving the right balance will require careful consideration of design development.

Volunteer demonstrations and targeting. There has been considerable discussion of the advantages and disadvantages of volunteer demonstrations. SSA demonstration authority necessarily requires volunteers and their written informed consent for the participants subject to the intervention (e.g., Stapleton et al. 2020). Some have argued that volunteer demonstrations are biased because they cannot replicate the impact of a permanent program that would impose the intervention on the

entire caseload. Others argue that volunteer demonstrations are superior because they permit estimation of impacts for those for whom the impact is likely to be the greatest, and that such evaluations at least provide an upper bound on general impacts. In their summary of implementation issues in SSA demonstrations, Wood and Goetz Engler (Chapter 9 in this volume) confirm that volunteers for work-incentive demonstrations are indeed those most inclined to work.

These discussions do not always recognize that the question is *not* what would happen if the entire caseload would take up the newly offered program were it offered to all of them. The question is whether—should the intervention be made permanent and offered to all—the fraction of those taking up the intervention would be the same as the fraction in the demonstration. If those who volunteer for a demonstration are reasonably similar to those who would take up the new intervention were it made national and permanent, then using volunteers should not be a problem.

Our view is based on our conclusion stated in the last section that it is quite likely that only a modest fraction of the SSDI or SSI caseload have enough residual work capacity to ever make major strides in employment, earnings, or program exit. With expectations lowered, concentrating attention on interventions that have effects on that small group of volunteers would seem warranted. However, we propose that interventions that are applied only to volunteers be conducted with the goal not of imposing the features on the entire caseload but only of offering them to a smaller subset of beneficiaries or recipients permanently. There would be no problem in principle of offering employment support programs, for example, which have a fixed number of slots and which are offered to volunteers or to those with identifiable characteristics that suggest higher probabilities of success. Such support programs could be implemented on a national level and on a permanent basis. However, it would be important that any demonstration aim to replicate such an intervention at the national level and be designed with that in mind.[57]

The idea of having a limited number of programmatic opportunities and encouraging, or even only allowing, those who appear to be most likely to succeed in them to receive them would require identifying those individuals. This raises the larger issue of targeting, which has been used relatively little tested in SSDI demonstrations.[58] There is a fairly large amount of evidence from existing demonstrations on heterogeneity in subgroup impacts based on differences in estimated program impacts by age, gender, education, past earnings, or impairment type, for example. The results often show no differential subgroup impacts and hence this body of work is inadequate to guide how targeting might be better accomplished. This is one area where much more research by SSA could be fruitful, with the goal to

[57] A rather different problem with volunteer demonstrations is the requirement that the volunteers be offered the opportunity to leave the demonstration and return to the current-law program. Some nationally imposed reforms might permit that, but some would not.

[58] See Weathers and Bailey (2014) for a discussion of targeting in the context of the Accelerated Benefit Demonstration.

better identify who is mostly likely to benefit from a reform and to concentrate the reform on that relatively modest number of SSDI beneficiaries and SSI recipients. The work of Maestas, Mullen, and Strand (2013), for example, suggests that residual work capacity can be estimated from outcomes of the application process, suggesting that the evaluations of work capacity arising from that process might help identify those most likely to respond to work incentives reforms. More investigation of the role of specific barriers to work (transportation, specific health barriers, etc.) is also needed; these are rarely included in measures of heterogeneity. The collection of more data on work histories of beneficiaries and recipients would be useful for the same reason. An additional object of research should be identifying work-related correlates of take-up in the first place, which is a necessary precondition to having an impact; heterogeneity in this respect has been little studied, yet the low rates of take-up which motivate this section suggest more research on this issue.

Estimating more than intent-to-treat impacts. In line with modern evaluation methodology, most demonstrations estimate only intent-to-treat (ITT) impacts, where evaluations estimate the impact of the offer rather than the impact of actual participation in the intervention. However, ITT impacts can be misleading if large shares of treatment group members do not take up the offered treatment. The low take-up rates in many demonstrations that we noted in the last section make this an issue of some empirical importance. It is likely that policymakers would like to know if a small ITT impact was the result of a small take-up rate but a large response by those who responded, or just a small response by the general beneficiary and recipient population. In the former case, investigation into the types of individuals who responded may also lead to targeting, as discussed above. We suggest that more demonstrations be designed with an aim to estimate additional treatment effects, such as the effect of the treatment on the treated, local average treatment effects, and marginal treatment effects if multiple scalable treatment groups were used, as noted above. As we have already suggested, based on the low take-up rates in the demonstrations we have reviewed and their modest ITT impacts, our view is that only a modest fraction of the SSDI and probably SSI caseloads have sufficient residual work capacity to achieve major increases in employment and earnings. The goal of estimating an average treatment effect on the treated, for example, is precisely the goal of estimating the impact of the treatment on those who take it up and then respond to it. More attention to these alternative treatment effects is consistent with the evidence we have reviewed above on small take-up rates and the likely modest number of beneficiaries and recipients who have major residual work capacity.[59]

---

[59] See Barnow and Greenberg (Chapter 2) and Weathers and Nichols (Chapter 3) for more detailed discussion. In some cases, defining who takes up the program is itself a difficult data and measurement problem because it requires comparing responses of members of the treatment group to members in the control group and defining what take-up and participation actually mean. But if the demonstration designs were to plan this in advance, data and measurement issues might be addressable at the design stage.

Assessing the effects of limited demonstration duration. An issue that we find to be underdiscussed in the evaluation reports we have read is the issue of limited duration of the demonstration. Participants know that the treatment is offered for only a finite period of time and that, if they do not exit SSDI or SSI before the demonstration ends, they will revert to the current-law program. Participants are quite likely to be affected by this knowledge and their response to the demonstration might be different than it would be if the intervention were implemented nationally on a permanent basis. This issue was considered in the Negative Income Tax experiments of the 1960s and 1970s, where theoretical analyses showed that the bias in an experiment of limited duration was ambiguous in sign, and could be favorable or unfavorable (Metcalf 1973). However, it is quite likely in the SSA demonstrations we have reviewed (e.g., POD) that the limited duration of the demonstration likely led to an underestimate of the response that would obtain in a permanent program, just because the demonstrations were not long enough for beneficiaries to understand and react to them.[60]

Entry effects. Entry effects of programmatic reforms in SSDI have been discussed for many years. Most have foundered on the problem that entry effects are very difficult to evaluate by RCT methods (Weathers and Nichols, Chapter 3 in this volume). However, it is also clear that the population of individuals considering applying are making decisions that are partly based on the relative financial and other attractions of applying versus not applying.[61] If nothing else, this is demonstrated by the fact that SSDI applications are sensitive to the business cycle and job availability. Given the inability to estimate entry effects credibly with RCT methods, we recommend that non-experimental designs be constructed to estimate entry effects, even if only estimating a range of possible such effects (see Chapter 3 for examples). These could then be combined with the estimated RCT effects of a programmatic reform on outcomes conditional on entry.

SUMMARY

In this chapter we have reviewed a number of demonstrations conducted by SSA on SSDI and SSI programs with an aim to increase employment and earnings among beneficiaries and recipients, and to increase exits from the program. Our review of 11 major demonstrations and other tests of SSDI and SSI programmatic reforms leads us to draw several lessons. One lesson we draw is that the impact of most demonstrations

---

[60] In one of the Negative Income Tax experiments, three different treatment groups were enrolled, each with a different duration of the offered treatment (SRI International 1983). The difference in response was then used to extrapolate to a permanent program.

[61] See Moffitt (1992a) and Hoynes and Moffitt (1999). Entry effects are unlikely to be important for small changes in SSDI work incentives rules, especially if they are poorly understood. But large changes in the experience of a large number of beneficiaries of the program could have a reputational effect that could affect entry.

on the already low rates of employment and exit are disappointingly small. Something of an exception occurs for some tests of reforms in the treatment of SSDI beneficiaries with mental impairments, although those programs are also quite expensive. We also reviewed one demonstration (AB) that intervened at the newly awarded stage that also had favorable effects but was quite expensive. But we also find that for most demonstrations, only a small number of individuals take up work programs offered in SSDI. This includes demonstrations with attempts to improve financial incentives, including the elimination of the cash cliff.[62] Implementation issues are important as are complexities in the interventions that are difficult for beneficiaries to understand.

Based on these conclusions, we suggest a number of ideas for future demonstrations worthy of further consideration. These include providing even stronger financial incentives to work through EITC-style programs that supplement earnings with benefits rather than reducing benefits; more testing of ways to reduce the work disincentives arising from the prospects of termination from the program; early intervention reforms such as those offering employment services to applicants at some stage of the application process; demonstrations to reduce work incentive rule complexity (either with or without offset); and partial SSDI benefit programs. We also suggest ideas for the demonstration design, such as possible increased use of scalable multiple treatment groups and factorial designs, more intentional planning of volunteer effects in demonstrations in conjunction with determining the effects of targeting, going beyond ITT effects in estimating the impact of program reforms, addressing the problem of limited demonstration duration, and incorporating entry effects.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Gubits for important assistance throughout the preparation of this chapter, Jeffrey Hemmeter for comments on an initial draft, and several SSA staff for comments on this draft. We also thank David Salkever and David Stapleton for helpful discussions.

[62] However, final results from the POD demonstration are not yet available and will be relevant.

Chapter 4

Comment

Hilary Hoynes
University of California, Berkeley

The chapter by Jesse Gregory and Robert Moffitt ("The Return to Work in Disability Programs") provides an excellent and insightful review of the literature on policies to promote returning to work among Social Security Income (SSI) recipients and Social Security Disability Insurance (SSDI) beneficiaries. This wide-ranging review includes a discussion of the conceptual issues around work supports and work incentives in this population, a detailed summary of SSA demonstrations to promote work, and suggestions for future reforms and demonstrations. They conclude that the results are disappointing; that is, there is little evidence that these policies have led to meaningful increases in employment and earnings.

In my comment, I would like to put the chapter and results in some context. I will focus on three main points: (1) why are we focused on work for this population? (2) an analysis of recent trends (pre-COVID) suggests some change is occurring; and (3) thoughts moving forward.

WHY ARE WE FOCUSED ON WORK IN THIS POPULATION?

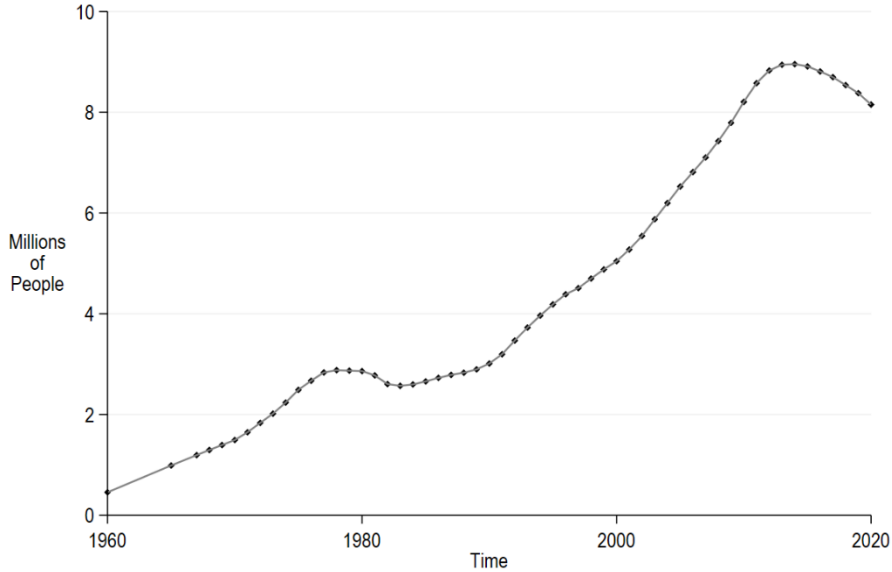
It is useful to start with some context around why there is so much policy attention on increasing work among SSDI beneficiaries and SSI recipients.

First, as mentioned by the authors, the reforms and demonstrations that they analyze have taken place during a time of large growth in SSDI caseloads and costs. As shown in Exhibit 4.2, the number of workers receiving SSDI increased from 3 million in 1990 to almost 9 million by 2010. Total program costs experienced similar increases during that time. As caseloads have increased, the composition of the SSDI caseload has changed. The share of the caseload with musculoskeletal diagnosis has increased from 20.6 percent in 1996 to 33.6 percent in 2019 (SSA 2020b).

Second, over this period there has been a steady decline in male labor force participation and more recently also a decline in female labor force participation. For example, the percentage of men age 16 and older in the labor force participation fell from 80 percent in 1970 to 75 percent in 1990 to below 70 percent in 2019. After rising for most of the 20th century, the percentage of women age 16 and older in the labor force peaked in the late 1990s and has steadily decreased since.[63] There is a large literature that explores these trends to identify the sources of the decline (see recent review by Abraham and Kearney 2020). Disability benefits seem to be part of the story, but there are other factors at play.

[63] See Figure 1 in Nunn, Parsons, and Shambaugh (2019) for recent data and discussion of trends in labor force participation.

Exhibit 4.2. SSDI Disabled Worker Beneficiaries, 1960–2019



Source: SSA Annual Statistical Supplement, 2019 (SSA 2020b).

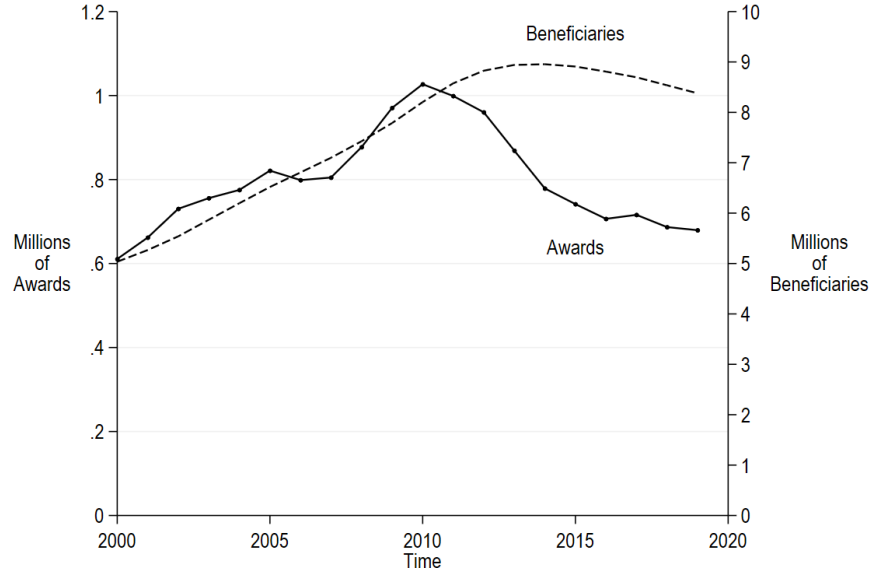
Taken together, the rising costs of the program alongside the declines in labor force participation naturally have led to policy discussions around how to increase employment among SSDI beneficiaries, with the goals of increasing aggregate labor in the economy and (possibly) reducing the costs of disability benefits.

BUT MORE RECENTLY THE TRENDS LOOK DIFFERENT

The discussion above relates to trends occurring over the past two decades or more. The most recent data (pre-COVID pandemic), however, show some important changes in trends. Exhibit 4.3 plots the number of beneficiaries (right axis) and awards (left axis) in SSDI between 2000 and 2019. New awards peaked in 2010 at just over one million and fell steeply to about 750,000 in 2015 before leveling off. Relatedly, and this has received less attention, the employment rate among disabled individuals is on an upswing after a many decade decline (see Exhibit 4.4, from Maestas 2019).

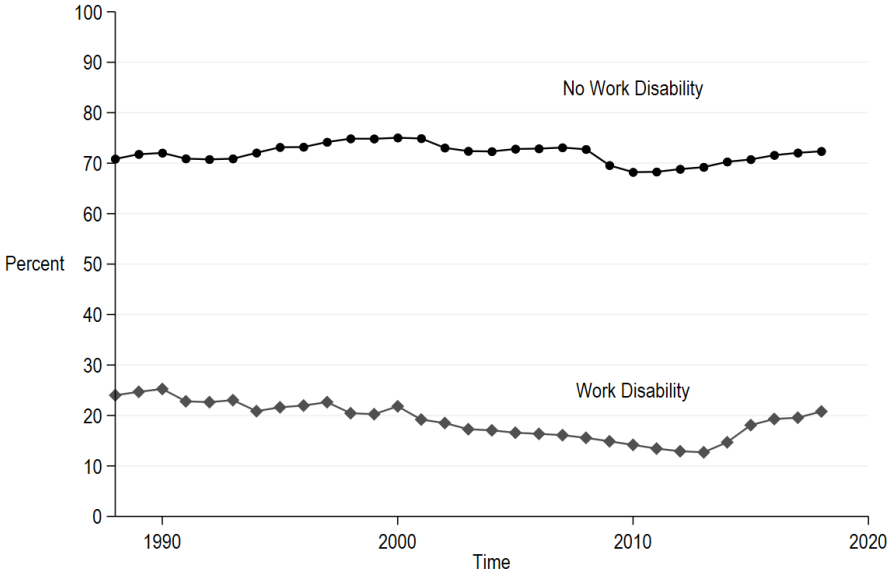
Maestas, Mullen, and Strand (forthcoming) find that the Great Recession, and the changes in labor market conditions, explain some of these recent trends in SSDI caseloads. However, changes in appellate-level decisions may also have played a role. Exhibit 4.5 presents trends in SSDI allowance rates at the hearing level between 1995 and 2018. Allowance rates have declined substantially in recent years—from almost 75 percent in 2005 to 50 percent in 2014. In ongoing work with Nicole Maestas and Alexi Strand, we examine the factors and policies that explain the reduction in allowance rates—with an interest in examining the effects on employment for persons with disabilities.

Exhibit 4.3. Number of Beneficiaries and Awards in SSDI, 2000–2019



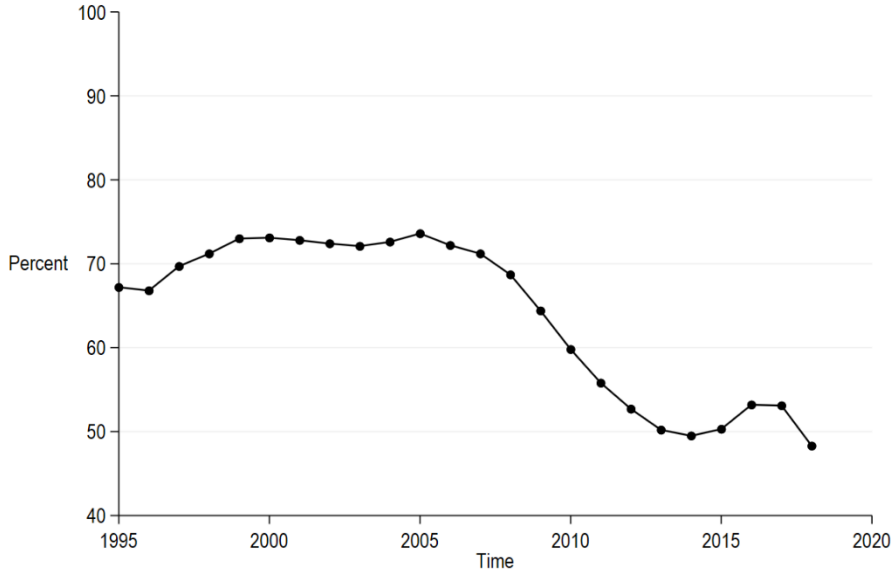
Source: SSA Annual Statistical Supplement, 2019 (SSA 2020b).

Exhibit 4.4. Employment Rate of People with Disabilities



Source: Data from Maestas (2019).

Exhibit 4.5. Allowance Rate at Hearing Level or Above, 1995–2018



Source: SSA Annual Statistical Supplement, 2019 (SSA 2020b).
Note: The *hearing level* is the level following reconsideration in the administrative review process. The hearing is a de novo procedure at which the claimant, the claimant's representative, or both may appear in person, submit new evidence, examine the evidence used in making the determination under review, give testimony, and present and question witnesses. The hearing is on the record but is informal and nonadversarial (SSA 2020b, Glossary).

THOUGHTS GOING FORWARD

First, it is important to put the goal of increasing employment and earnings among disabled works into some perspective. The standard public finance framing for considering the "optimal" design of transfer programs is to examine the tradeoffs of protection versus distortion. For SSDI, the goal is to provide protection against disability-related earnings losses but balanced against not inducing labor force nonparticipation among people who could otherwise work. To make sense of the results here (policies and demonstrations around work and work disincentives) we need to also know about protection.

We need more evidence on the effects of SSDI (and SSI) on short- and long-term health, economic and financial well-being (e.g., Deshpande [2016a] on SSI).

Second, we need to build on the encouraging evidence reviewed by Gregory and Moffitt showing that early interventions and health care coverage may increase work among the disabled. This suggests we need to identify ways to keep more SSDI applicants in the labor force—rather than try to affect their behavior once they are receiving disability program benefits. A multi-tiered system might be a good structure. One tier would target those with capacity to work (partial insurance, as promoted by

Maestas [2019]). Another tier would provide health care and short-term income supplements. The third tier would target those with long-term disability without capacity to work.

Chapter 4

Comment

Kathleen Romig
Center on Budget and Policy Priorities

In their chapter ("The Return to Work in Disability Programs"), Gregory and Moffitt provide an excellent overview of several decades of the Social Security Administration's (SSA's) return to work demonstrations. Their findings may surprise and discourage some readers who expected these experiments to increase work for a substantial number of beneficiaries and recipients—especially if they were hoping increased work would increase program exits and thereby reduce costs. For example, Gregory and Moffitt find that most SSA work demonstrations do not meaningfully increase employment, earnings, or labor force participation. Even when beneficiaries' and recipients' earnings increase, they rarely rise above the Substantial Gainful Activity (SGA) level. What is more, there are "essentially never" increases in program exits due to work from demonstrations, and "rarely" reductions in Social Security Disability Insurance (SSDI) program expenditures. Only a small number of SSDI beneficiaries try the interventions offered in work demonstrations—possibly, the authors posit, because few have residual work capacity.

Given the very strict medical and vocational criteria for SSDI and Supplemental Security Income (SSI) program benefits, should the fact that few beneficiaries have residual work capacity surprise readers? Consider how the Social Security Act defines a qualifying disability:

> The term "disability" means inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months. (42 USC § 423(d)(1)(A))

Furthermore, it must be

> of such severity that he is not only unable to do his previous work but cannot, considering his age, education, and work experience, engage in any other kind of substantial gainful work which exists in the national economy, regardless of whether such work exists in the immediate area in which he lives, or whether a specific job vacancy exists for him, or whether he would be hired if he applied for work. (42 USC § 423(d)(2)(A))

Perhaps the fact that few SSDI beneficiaries and SSI recipients are able to work at SGA over a sustained period is *not* evidence that the demonstrations have failed,

but evidence that the disability determination process has succeeded in identifying applicants who meet these extraordinarily strict criteria.

In fact, in addition to their work-limiting disabilities, SSDI beneficiaries and SSI recipients face many other barriers to work.

- Disability beneficiaries have lower education levels than average. Disability can limit a person's education, and limited education is also correlated with higher levels of disability. Limited education constrains employment opportunities—especially for workers whose disabilities prevent physical work.
- Disability beneficiaries are older than average. The older a person gets, the more likely he or she will acquire a disability. Advanced age poses another employment barrier; it is more difficult for older workers to find jobs or to shift to other kinds of work.
- Disability beneficiaries typically have substantial time out of the labor force. Most SSDI applicants have been out of work at least a year before they even apply for benefits, and for SSI disability, work history tends to be even more limited. The application process for both programs takes months—or sometimes years, if appealed. During this time, workers' skills and connections atrophy, and gaps in employment history pose an additional employment barrier.
- Finally, people with disabilities face structural barriers to employment. Disability discrimination is pervasive. Many employers do not offer sufficient accommodations to allow disabled employees to work. Transportation can pose another barrier; public transit is often unavailable, inaccessible, or unreliable. And many would-be workers with disabilities do not have access to the health care they need, particularly the long-term services and supports that would allow for employment.

Given that all disability beneficiaries have work-limiting disabilities and most face additional barriers, it is little wonder few SSDI beneficiaries or SSI recipients are able to perform substantial, sustained work, even with incentives to do so.

A DIFFERENT WAY OF MEASURING WORK SUCCESS FOR DISABILITY BENEFICIARIES

If sustained, substantial work is not possible for most disability beneficiaries, is it possible to construct a successful return to work program? The answer to that question depends on how one measures success.

If policymakers expect most disability beneficiaries to return to work, success isn't likely, past demonstrations show. Instead, policymakers could aim to reduce barriers for those with the capacity and desire to work. Gregory and Moffitt suggest that SSA focus employment interventions on volunteers who seem most likely to succeed. This would allow the agency to tailor interventions to the needs of SSDI

beneficiaries and SSI recipients with specific characteristics and improve their odds of success. Evidence supports this approach, as Gregory and Moffitt point out—SSA's mental impairment-based demonstrations resulted in rare employment and earnings increases.

If policymakers aim to reduce SSDI and SSI program spending, success isn't likely, either, based on the evidence. Instead, policymakers could aim to improve beneficiaries' and recipients' well-being overall. Many people with disabilities who work report benefits beyond their wages—improvements in mental health, community integration, and a sense of purpose and connection. But these effects are often not measured in demonstration projects. When they are—for example, the Mental Health Treatment Study measured improvement in mental health—they show work can have positive effects beyond earnings. The ongoing Supported Employment Demonstration is another example of a study that measures mental health and quality of life, in addition to financial outcomes.

Shifting from broad-based interventions aimed at reducing costs toward focused interventions aimed at improving quality of life would require not only different services and metrics, but also likely more money, as Gregory and Moffit point out. If policymakers' goal is truly to encourage work among disability beneficiaries, it would be money well spent. But if the goal is to save money by reducing enrollment, it seems likely that another round of work demonstrations will only bring more surprise and disappointment.

Chapter 5

Demonstration Evidence of Early Intervention Policies and Practices

Kevin Hollenbeck
W.E. Upjohn Institute for Employment Research

Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) are important components of our nation's social safety net for individuals with disabilities. They provide benefits for individuals whose disability precludes them from (substantial) gainful employment. However, if policies or practices for individuals who have a disability or who experience a disabling injury or illness can be implemented effectively that allow them to maintain or to achieve meaningful employment and earnings and forgo applying for benefits, then both they and taxpayers will benefit—a win-win situation! These policies and practices aimed at keeping individuals from entering the SSDI or SSI programs prematurely include transition assistance for youth (ages 16–24) and "early intervention" policies or practices targeting the population of adults (ages 25–64) with disabilities. This chapter focuses on the latter.

The Social Security Administration (SSA), which administers SSDI and SSI, would benefit from an effective early intervention, but it has limited ability to implement one. As with any agency, its goal is to be as efficient as possible; that is, to use its limited resources to provide benefits to individuals with disabilities. It should be noted that other agencies provide training, education, or other support services to these individuals.

In 2019, SSA received approximately 2 million applications for SSDI from workers and about 1.3 million applications for SSI from individuals ages 18–64. However, only around one-third of the SSDI and SSI applicants have been or will be approved for benefits.[1] If an early intervention were implemented that would effectively stem the inflow of applications, especially those likely to be denied, SSA could save costs: reduced operational costs from processing fewer applications, as well as reduced benefit payments. An enigma for SSA, however, is that it has very limited interaction with individuals who experience a disabling event until they apply for benefits. So SSA has limited opportunity to affect the behavior of individuals prior to their applying for benefits for the first time.

Besides attempting to reduce the inflow of first-time applications, SSA could find it beneficial to facilitate the return to work of individuals whose applications for

[1] The final award rate for disabled-worker applicants has varied over time, averaging 32 percent for claims filed from 2009 through 2018 (SSA 2020b, 155).

benefits have been denied.[2] If individuals do return to work, revenue from the payroll tax will increase, and the likelihood of later, repeated applications would likely decrease. Again, however, the ability of SSA to affect these individuals is limited as they are neither receiving benefits nor actively pursuing an application.

If effective early interventions were to stem the inflow of new or repeated applications that are likely to be denied, then another group of individuals who would benefit is future eligible applicants. These individuals would likely receive more timely decisions.

Despite limitations in being able to interact with individuals before they apply for benefits or after their application has been denied, SSA has used and is using its demonstration authority to test early intervention initiatives, often in collaboration with other agencies. This chapter provides analyses of the evidence that has been or is being generated by these demonstrations and other ideas that have been implemented or put forward for early intervention programs or policies.

To be successful, early interventions will increase the potential applicant's productivity through supports that will enhance their human capital. The early intervention demonstrations or reforms described in this chapter are intended to increase individuals' productivities and effective wage rates through the provision of Vocational Rehabilitation (VR) services or other types of training, through provision of assistive technologies or other work accommodations, by intervening as soon as possible after the medical event, by providing the services of a workforce counselor, through transitional jobs that will increase worker human capital, or by standardizing the eligibility determination process making it less manipulable to varying diagnoses from local/personal medical staff.

The first section of this chapter describes the context for early intervention strategies and characterizes the target populations for such strategies. The second section of the chapter is the heart of the discussion. It reviews the SSA demonstrations and reforms that have taken place. The review places particular emphasis on the empirical evidence that has been gathered to date. The third and fourth sections, respectively, complement the review of demonstrations by discussing international experiences and by presenting early intervention strategies suggested in the literature that have not been tested or implemented. Those sections are followed by a presentation of the lessons about early intervention policies and practices that can be drawn from the empirical evidence or suggestions in the literature. The sixth section offers an idea about a potential demonstration of an early intervention targeted at individuals age 50 and older that SSA might consider implementing. The final section offers concluding remarks.

---

[2] A focus of this chapter is on denials, but effective early interventions may reduce the inflow of (approved) beneficiaries, as well. Maestas, Mullen, and Strand (2013) estimate that 18 percent of new SSDI beneficiaries are able to engage in substantial gainful activity within two years, but only 5 percent do.

CONTEXT

Individuals' Pathways to Benefit Application or Employment

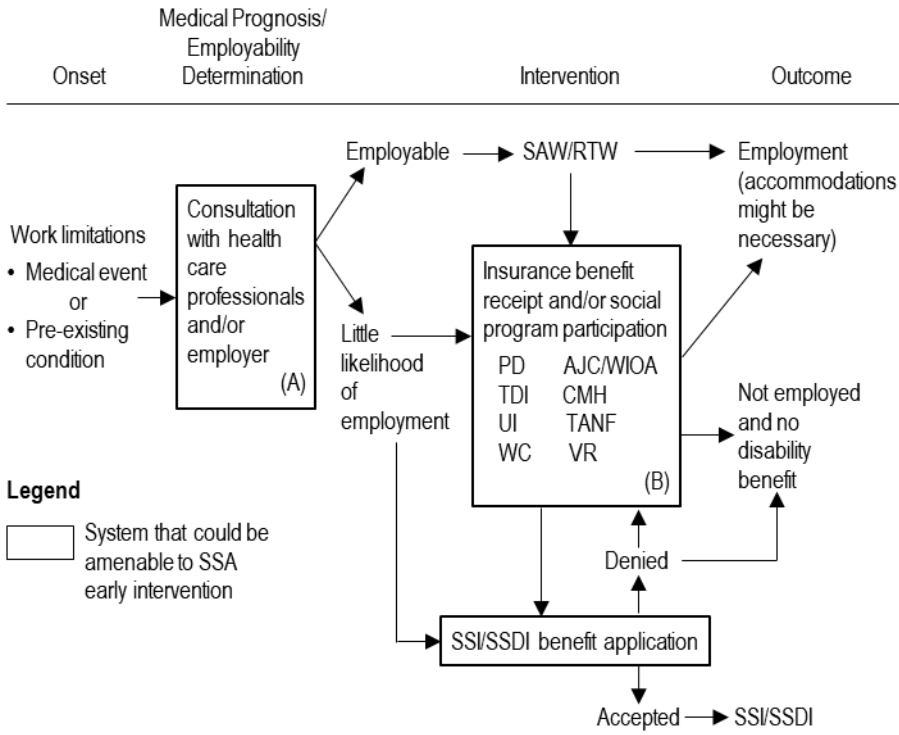
Though the circumstances of individuals with disabilities can vary considerably, the decision to apply for disability benefits (SSDI or SSI) essentially involves four steps, as shown in Exhibit 5.1 below. The steps are onset, medical prognosis and employability determination, intervention, and outcome. The onset, as displayed in Exhibit 5.1, occurs when an individual experiences a job-threatening injury or illness or when they decide that a medical condition that they have had for a long time affects their ability to work. Then, in consultation with one or more medical professionals and/or their employer (box A), the individual will learn of their medical prognosis, from which they will be able to ascertain their likely employability. In the US health care system, this step is highly decentralized and individual driven. Some providers might encourage employment; others might be more circumspect. Based on the information the individual receives, they will perceive themself as (or might be told that they are) employable or not likely to be employable.

Individuals who perceive themselves to be employable and who are interested in obtaining or retaining employment will seek stay-at-work (SAW) or return-to-work (RTW) interventions. In some cases, the individual and their employer will agree to continued employment with accommodations such as job redesign, reduced time, assistive devices, or other practices that allow the individual to continue to be able to productively do their job. In most other cases, in which job separation has or will occur, interventions will usually involve an insurer that will provide short-term financial assistance (left column of box B in the exhibit) and a social program that will provide assistance such as financial aid, occupational training, job search services, or other types of support aimed at the goal of re-employment (right column of box B). These individuals compose the target population for early interventions.

Individuals who perceive themselves as having little likelihood of re-employment might receive insurance benefits and participate in social programs, but they are likely to also successfully apply for disability benefits, especially if their condition is on the SSA Listing of Impairments–Adult Listings (Part A).[3] It is possible that these individuals become gainfully employed, but because the likelihood is low, they are excluded from the target population for early intervention interventions in the exhibit.

[3] The listing of impairments for adults is available at https://www.ssa.gov/disability/professionals/bluebook/AdultListings.htm.

Exhibit 5.1. Pathway to Benefit Application



Key: AJC/WIOA=American Job Centers/Workforce Innovation and Opportunity Act agency. CMH=community mental health. PDI=private disability insurance. RTW=Return to Work. SAW=Stay at Work. TANF=Temporary Assistance for Needy Families. TDI=state-mandated temporary disability insurance. UI=Unemployment Insurance. VR=Vocational Rehabilitation. WC=workers' compensation.

The possible outcomes in the exhibit are employment, in which accommodations might be necessary, neither employed nor on benefits, and receipt of SSDI/SSI. The outcomes are not mutually exclusive. The programs and supports that SSA provides to SSDI beneficiaries and SSI recipients to encourage employment, such as the Ticket to Work program, result in some individuals whose outcomes are both employment and benefit receipt.

Exhibit 5.1 depicts the two systems along the pathways to benefit application that could be amenable to an SSA early intervention strategy—boxes A and B. What is striking is that both of these systems are extremely decentralized, which makes any sort of SSA collaboration extremely difficult. Box A, in which the Retaining Employment and Talent after Injury/Illness Network (RETAIN) demonstration described below is encapsulated, comprises thousands of health care professionals in independent practices. Box B comprises the social assistance system described in some detail below, with several programs that are operated independently and that have state

and local autonomy. The Supported Employment Demonstration (SED) described below is a collaboration between SSA and one of the agencies within this second system.

Size and Characteristics of the Target Populations

Size

Two independent methods have been employed for estimating the size of the target population for whom early intervention(s) could be effective. First, the number of adults ages 25–64 who experience a serious on- or off-the-job injury or illness in a year is estimated.[4] Second, the number of SSDI or SSI applicants in a year who were denied benefits because they were determined to be employable is estimated.[5] These estimates suggest that the target population for early interventions was between 1.1 and 2.0 million individuals in 2019.

The starting point for the initial estimate of the flow of individuals for whom early interventions might be targeted is individuals who experience a disabling medical event—injury or illness—on the job. The Bureau of Labor Statistics (BLS 2020b) reported that approximately 890,000 nonfatal work-related injuries and illnesses that resulted in the loss of at least one day of work occurred in private industry and approximately 220,000 in state and local government enterprises in 2019. The BLS provides worker characteristics and number of days away from work for the private industry occurrences. These data indicate that approximately 220,000 cases resulted in 31 or more days away from work for workers ages 25–64. Assuming that the severity and characteristics of the injuries and illnesses for workers in state and local government enterprises are similar to those in private industry adds approximately 50,000 cases. The BLS data do not include the federal workforce, however. According to Hill (2020), the civilian federal workforce is slightly less than 20 percent the size of the state and local workforce, so federal workers might add another 10,000 cases. In short, approximately 280,000 adult workers in the United States suffered an on-the-job injury or illness in 2019 for which they missed at least 31 days of work.

To round out the estimate, non-occupational injuries or illnesses are added. According to the National Safety Council (2020), more than three times as many injuries requiring medical attention occur off the job than on the job. Indeed,

---

[4] This estimate undercounts the target population because it does not include the number of individuals who have not applied for SSDI but who have congenital or other pre-existing disabilities or illnesses that occurred in prior years who might be employable. On the other hand, it overcounts the number of individuals experiencing a serious disability or illness in a given year who might apply for benefits, as SSA rules require the medical condition to last or be expected to last 12 months or end in death, and very little data exist on the duration of injuries or illness that occur in a year.

[5] SSA could presumably generate the precise number, but it is not easily producible from public data.

Neuhauser (2016) estimates that non-occupational injuries and illnesses could be six times as common as occupational ones. If these factors are applied to the on-the-job lost-time injury data, then an estimate of 1.1 to 2.0 million individuals who experienced a work-threatening medical event in 2019 either on or off the job is derived.

As a check on that estimate, the number of individuals who apply for SSDI or SSI but who have been or will be denied because of a determination that they are employable is estimated. Whereas around 3.3 million applications for SSDI or SSI were received from adults in 2019, many individuals applied for both. Of these, approximately 2.4 million adults applied for one or for both programs in 2019.[6] With an ultimate approval rate of 35 percent[7] and under the assumption that the 2.4 million is a steady-state estimate of individuals who apply for benefits, then it can be estimated that approximately 840,000 of the adults who experienced a disability in 2019 will ultimately be approved for SSDI or SSI. (Note that in 2018, there were about 500,000 SSDI allowances and 460,000 SSI awards to individuals ages 18–64, so an estimate of 840,000 adults seems reasonable.) The share of the SSDI or SSI applications that will be ultimately denied is about 65 percent, or about 1.56 million adults.

Data presented in Wixon and Strand (2013) show that approximately two-thirds of applications for SSDI or SSI were denied because the individuals were determined to be capable of work.[8] Assuming that two-thirds also holds for individuals who applied for benefits in 2019, then the size of the population of individuals who have been or will be denied benefits who are capable of working is about 1.0 million individuals. Maestas, Mullen, and Strand (2013) estimate that 18 percent of new SSDI beneficiaries are able to engage in Substantial Gainful Activity (SGA) within two years, but only 5 percent do. If 18 percent of the allowed beneficiaries in 2019 were able to engage in SGA within two years, then approximately 130,000 individuals should be added to the number of individuals denied benefits due to a capability of working, in order to finalize the second estimate of the annual population who could

---

6 Computed from data provided at "SSA State Agency Monthly Workload Data," https://www.ssa.gov/disability/data/ssa-sa-mowl.htm (accessed May 7, 2021).

7 Ultimate award rates averaged 34.2 percent between 2009 to 2015 (SSA 2020b, chart 11; accessed May 7, 2012, at https://www.ssa.gov/policy/docs/statcomps/di_asr/2019/charts-text.html#chart11). In the years 2012 to 2015, the data show less than 1.0 percent with the final decision pending. Years after 2015 had at least 2.4 percent pending.

8 Wixon and Strand (2013) display the Disability Determination Services codes for applications to SSDI (Table 1) and SSI (Table 2). For the former, 67.0 percent of the denials have codes of H1, H2, G1, G2, J1, or J2, which denote capacity for SGA. For SSI applicants, 70.5 percent of the denials have codes of 31, 42, 32, or 43, which denote capacity for SGA. In an analysis of the individuals eligible for the SED demonstration (denied SSDI or SSI applicants, as discussed below), Taylor et al. (2020) report that the mean of their WORKPOTENTIAL variable, which is defined as having a denial code of N32, which is ability to earn above the SGA level in any job, is 37.2 percent (Table 6-5), which compares to 43.3 percent for code 32 in Table 2 of Wixon and Strand (2013).

benefit from SAW/RTW supports; that is, a bottom line of approximately 1.2 million individuals.[9]

Thus, two independent estimates suggest that the target population is between 1.1 and 2.0 million individuals. This estimate accords with the discussion presented by Hollenbeck (2015). Using estimates from the Census Bureau's Survey of Income and Program Participation and its Current Population Survey, he reports that more than 2 million individuals annually leave the labor force, at least temporarily, due to a disability.

Characteristics

The above discussion derives an estimate of the size of the target population, but very little data are available to characterize it. BLS (2020b) provides selected characteristics for private industry workers who experienced an on-the-job injury or illness that caused at least 31 days of work. Men, workers older than age 45, workers of color, and workers engaged in a few unskilled or semi-skilled occupations are overrepresented. It should be noted that it is not clear how representative the characteristics of individuals with work-related impairments will be of the full target population. Hollenbeck's study finds similar characteristics for individuals who left the labor force because of a disability, except for gender:

> Individuals who exit the labor force because of a disability are disproportionately female and non-white; they are less likely to have attended college and their average age is over 45. In particular, among the individuals who reported disability onset at some point in the eight months' worth of [Current Population Survey] data we examined, about 20 percent did not have a high school diploma, and only about 22 percent had a bachelor's degree or higher; in the labor force overall, these percentages were about 10 percent and 33 percent, respectively. (2015, 3)

As Exhibit 5.1 above notes, individuals with disabilities participate in a number of different social insurance or social assistance programs prior to or coincident with applying for disability benefits from SSA. Any early intervention project or policy that gets implemented is likely to interact with one or more of these policies or programs. To frame the context for early intervention policies, the next subsection briefly describes the history of those programs and denotes their interactions with SSA.

---

[9] Autor et al. (2015) suggest that the percentage of beneficiaries able to engage in SGA is considerably underestimated in Maestas, Mullen, and Strand's (2013) study because it took into account only the effect of receiving benefits on recipients' labor supply, whereas there is an additional impact caused by the delay in processing. This finding suggests that the number of 2019 beneficiaries capable of earning SGA might be twice as large as the estimate of 130,000, making the bottom line estimate around 1.5 million.

US Policies and Programs for Disabled Adults

This subsection describes the policy and programmatic environment for disabled adults in the United States within which the SSDI and SSI programs operate.[10] It is likely that successful early interventions will involve collaborative efforts with some of these programs.

Workers' Compensation (WC)

In 1911, Wisconsin was the first state to mandate that employers maintain workers' compensation insurance. Gradually, WC became required in all states.[11] Although there is wide cross-state variation in the program's regulations and benefits, the basic tenet underlying this insurance is that workers who are injured, or in some states who become ill, while on the job would have their medical expenses paid and would receive compensation for loss of time, in return for waiving the right to sue their employer. This insurance provides an important economic safety net for a subset of disabled adults—those whose disability occurred in the workplace. Most workplace injuries do not involve lost time; but if a worker does lose time, WC will include benefit payments, which in most cases are time limited. In general terms, the earnings replacement rate for WC is around 60 percent.

SSDI interacts with WC in two ways. SSDI beneficiaries might have received WC prior to receiving benefits from SSA; and in some cases, individuals are receiving both. Nichols et al. (2020) find that WC can be a "touchpoint" for individuals with disabilities prior to receiving SSDI or SSI benefits, although the percentages are relatively small. In their analysis, 10 percent of a national sample of nonelderly individuals with a work disability received WC, and 19 percent of them received disability benefits within 20 months of their earnings loss. O'Leary et al. (2012) analyzed a comprehensive data set of WC claims matched to SSA administrative data from one state. They report that about 10 percent of individuals with a compensable lost-time injury received SSDI within 10 years of the injury, a major share of whom were permanent total disability cases. The Social Security Act was amended in 1965 to require an offset in WC or SSDI benefits if they were being received at the same time and exceeded a threshold.

Given the decentralized nature of funding for and rules and regulations for WC, any collaborative effort by SSA to implement an early intervention strategy would likely require interactions with all 50 states and the District of Columbia.

---

[10] This discussion is limited to civilian benefits. For a discussion of benefits available to veterans, see "VA Disability Compensation," US Department of Veteran Affairs (http://www.va.gov/disability/) and further links on that page.

[11] Texas does not require private sector employers to provide WC, although public sector employers are required to carry it. Private employers may buy voluntary WC insurance.

Vocational Rehabilitation (VR)

VR provides services—training, education, counseling, and so on—that will enable individuals with disabilities to obtain employment. Not long after Wisconsin enacted the first state WC law, the Smith-Fess Act of 1920 established the VR program for individuals with physical disabilities. A little more than two decades after the original act, in 1943, the Barden-Lafollette Act expanded VR eligibility to individuals with mental impairments or psychiatric issues. The Rehabilitation Act Amendments of 1992 emphasized presumptive employability—individuals with disabilities should be assumed to be employable unless proven otherwise—as the primary goal of VR.

Individuals with disabilities who are interested in employment are eligible for VR.[12] Depending on the individual's circumstances, cost sharing could be required from the participant. The services that are provided are wide ranging: medical and psychological assessment, vocational evaluation and planning, career counseling and guidance, training and education after high school, job-site assessment and accommodations, job placement, job coaching, on-the-job training, supported employment, assistive technology and devices, and time-limited medical and/or psychological treatment.

It should be noted that under the VR Cost Reimbursement program, SSA reimburses VR agencies for the service costs of SSI recipients or SSDI beneficiaries who become employed for at least nine months and earn at least the level of SGA (currently $1,310 per month for disabled individuals and $2,190 per month for individuals who are blind). SSA reimbursed the training costs for almost 18,000 individuals in FY 2020 (SSA, "State," n.d.).[13]

Community Mental Health

The Community Mental Health Centers Act of 1963 was an initiative of President John Kennedy. The main focuses of the Act were to provide medical and other services to individuals with long-term and disabling illnesses and to deinstitutionalize individuals with mental impairments. At the federal level, funding for community mental health agencies comes from the Substance Abuse and Mental Health Services Administration within the US Department of Health and Human Services. States provide supplemental funding. Though much of the emphasis of the community agencies is on treatment, supported employment and employment counseling services are also provided. As noted below, a large-scale early intervention demonstration

---

[12] The severity of the disability will affect eligibility in the state agencies under "order of selection." These states must prioritize individuals with the most severe disabilities.

[13] In fact, state VR agencies are the only programs authorized by law and automatically approved to provide services to beneficiaries under the Ticket to Work program without becoming an Employment Network. Whenever a beneficiary receives services from a state VR agency, SSA considers the individual to be using their ticket (https://www.ssa.gov/work/vocational_rehab.html, accessed May 7, 2021).

(SED) is being conducted in collaboration with community mental health agencies. Given the federal funding and administrative base of the community mental health program, it would seem possible for SSA to implement a training cost reimbursement intervention similar to the one it has with VR. However, employment training that results in earnings that exceed the SGA level might be less of a primary goal of these agencies than in VR. For example, the Mental Health Treatment Study (MHTS) found that the intervention that was tested improved the labor market and mental health outcomes for SSDI beneficiaries but had virtually no effect on the number of individuals with earnings that exceeded the SGA level.[14]

Private Disability Insurance (PDI)

Short- and long-term disability insurance plans, typically provided and financed by employers, replace some of the pay workers lose when they cannot work because of an injury or illness that is not related to their job. According to the BLS (2020a), 42 percent of private industry workers had access to short-term disability insurance plans and 34 percent to long-term plans. Some 26 percent of state and local government workers had access to short-term and 38 percent to long-term coverage. Long-term disability benefits are often set at 60 percent of prior earnings, and they are paid until the individual recovers, until retirement, or for a specified number of months. Benefits are typically coordinated with SSDI benefits. For example, plans may require beneficiaries to apply for SSDI and may reduce the benefit by some or all of the SSDI benefit received.

Autor and Duggan (2010) suggest that legislatively mandating PDI coverage could be an effective early intervention. Nichols et al. (2020) find that as with WC (discussed above), PDI can be a "touchpoint" for disabled individuals with disabilities prior to receiving SSDI or SSI benefits, although the percentages are relatively small. In their analysis, 9 percent of a national sample of nonelderly persons with a work disability received PDI, and 37 percent of those individuals received disability benefits within 20 months of their earnings loss.

[14] The following comes from the MHTS final report (Frey et al. 2011): "Eight percent of the study participants showed average earnings over the 24-month study period that exceeded the current level of SGA. Beneficiaries in the treatment group did not experience an increase in work that SSA considers SGA when compared to participants in the control group. Neither did participants in the treatment group experience a reduction in benefit payments when compared to participants in the control group" (EX-9).

State-Mandated Temporary Disability Insurance (TDI)

Five states mandate insurance programs that partially compensate workers for the loss of wages caused by a temporary disability not related to their job.[15] In 1946, Congress amended the Federal Unemployment Tax Act to permit states where employees make contributions to the Unemployment Insurance (UI) program to use some or all of these contributions to pay disability benefits (but not administrative costs). Rhode Island had already enacted this program in 1942. It was followed by California (1946), New Jersey (1948), New York (1949), and Hawaii (1969). Rhode Island has an exclusive state fund; in the other states, employers may buy group insurance or self-insure. The plans vary across states, but in general, benefits are approximately 50 percent of earnings and have maximum durations of 26 or 52 weeks.

Other Programs

Programs that serve sizeable numbers of adults with disabilities but do not directly target that population include Temporary Assistance for Needy Families (TANF), workforce development programs operated under Title I of the Workforce Innovation and Opportunity Act (WIOA), UI, and the Supplemental Nutrition Assistance Program. Barden (2013) notes that depending on the definition of disability that is used, between 10 and 40 percent of TANF recipients could be characterized as having a disability. WIOA performance data report that in Program Year 2019 approximately 34,000 exiters (6.4 percent) from Title I WIOA Adult or Dislocated Worker programs were individuals with disabilities (DOL, "WIOA," n.d.). According to BLS (2019), about 61,000 of the 947,000 individuals (6.4 percent) who received UI benefits in 2018 self-reported a disability. In 2019, almost 12 percent of Supplemental Nutrition Assistance Program benefits went to households with an individual with a disability.

EARLY INTERVENTION DEMONSTRATIONS

This section of the chapter will review the early intervention demonstrations that have been administered or supported by SSA. The discussion proceeds in chronological order.

---

[15] Recently, two states (MA, WA) and the District of Columbia passed family and medical leave mandates. Not targeting solely individuals with disabilities, these programs include coverage for temporary disabilities. The Massachusetts program is described at https://www.mass.gov/info-details/paid-family-and-medical-leave-pfml-fact-sheet (accessed May 12, 2021); Washington State at https://paidleave.wa.gov/ (accessed May 12, 2021); and the District of Columbia at https://does.dc.gov/page/dc-paid-family-leave (accessed May 12 2021).

Demonstration to Maintain Independence and Employment (DMIE)

Authorized under the Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act), the purpose of DMIE was to see whether early intervention of medical assistance and employment supports could delay or prevent reliance on SSDI or SSI and loss of employment. The Centers for Medicare and Medicaid Services provided funding to Medicaid agencies in Texas, Minnesota, Kansas, and Hawaii to develop, implement, and evaluate interventions under DMIE.

Whalen et al. (2012) report on an experimental evaluation of interventions developed under DMIE in those four states that served individuals between 2006 and 2009. Eligibility requirements for individuals to participate in the demonstrations were ages 18–62, working at least 40 hours per month, and not receiving SSI or SSDI. The states varied widely with respect to the medical conditions targeted and to the medical benefits provided to the treatment group participants. Texas targeted individuals with behavioral issues and provided enhanced and expedited mental health services. Minnesota also targeted individuals with behavioral health issues and provided medical transportation and a health club membership. Kansas enrolled individuals with a wide range of physical and mental conditions from a high-risk insurance pool. Enrollees from Kansas were provided with physical therapy and home health visits. Hawaii enrolled workers with diabetes and provided medical therapy management, diabetes education, and nutrition counseling. Three of the four states provided career counseling to treatment participants (Kansas was the exception).

The evaluations analyzed three sets of outcomes: health and functional outcomes, employment and earnings, and application and receipt of SSDI or SSI. Two limitations of the evaluation should be kept in mind in interpreting the findings. First, the follow-up period was only 24 months, which is arguably a short time frame to observe substantial changes. Second, the analytical sample sizes were quite modest, so outcome differences between treatment and control groups had to be fairly sizeable to be statistically significant. The outcomes that were measured with SSA administrative data—earnings and SSDI/SSI application and receipt—have the largest analytical samples whereas the sample sizes for the other outcomes were reduced substantially because of missing data. Within the SSA administrative data, Texas had the largest analysis sample—approximately 900 in the treatment group and about 700 in the control group. Minnesota had approximately 900 in the treatment group and about 270 in the control group.

Among health and functional outcomes, the DMIE interventions improved physical health and reduced functional activity limitations only in Hawaii (with undetectable impacts in the other states). In Minnesota, mental health improved. No states saw changes in any of the employment outcomes, which is not too surprising given that the treatments were primarily of a medical nature and an eligibility criterion was being employed at least 40 hours a month. The latter implies that the individuals receiving the DMIE treatments had jobs, so any employment impact would have been

through number of hours worked or through job changes, which seem unlikely given the focus of the DMIE interventions.[16]

The DMIE interventions lowered disability benefit applications and receipt in Texas, but did not have an impact in the other three states.[17] About 12 percent of the treatment group participants in Texas applied for SSDI or SSI within two years after DMIE enrollment, compared to 14.5 percent for controls. This 2.5 percentage point impact represents a relative impact of about 17 percent, weakly statistically significant at the 80 percent level. When the percentages of individuals who received benefits within one year were analyzed, the reduction in Texas SSI recipients from 3.3 to 1.7 percent was statistically significant (at the 95 percent level). The reduction in SSDI recipients was not statistically significant, however.

TANF-SSI Disability Transition Project (TSDTP)

Launched in October 2008 in collaboration with the Administration for Children and Families (ACF) within the US Department of Health and Human Services, the TSDTP attempted to gauge the extent to which individuals with disabilities who were receiving TANF were also receiving SSI or were planning to apply for SSI. As noted above, TANF serves many families with individuals with disabilities; in some instances, the disabilities exempt TANF recipients from its work requirements. From an early intervention perspective, the project shed light both on whether TANF was a gateway to SSI and on whether an intervention could be offered that enhanced the employability of TANF recipients, thereby decreasing the likelihood of their applying for SSI.

The project comprised two phases: in the first phase, administrative data were analyzed and field visits were conducted in seven sites across five states; in the second phase, programmatic interventions were pilot-tested in three counties. The analyses conducted in the first phase (Farrell and Walter 2013) concluded the following:

- Only a small percentage of SSI applicants had received TANF in the year prior to application.
- TANF staff were, in general, not familiar with the SSA disability determination process.
- TANF recipients who applied for SSI were equally likely to be awarded benefits as were SSI applicants not receiving TANF.

---

[16] Whalen et al. (2012) note a limitation in their measure of employment: "Using positive hours worked as a measure of employment may overestimate the number of people who are employed. Based on this definition, a person working only one hour a week is considered to be employed" (46).

[17] Whalen et al. (2012) present disability application and receipt outcomes for the joint sample of Texas and Minnesota, but the results are mainly driven by the Texas impacts.

These conclusions suggest that at least at the time of the project, TANF was not a gateway that led to a large number of SSI applications or participants.

In the second phase of the project (Farrell et al. 2013), interventions were assessed in three different counties: Ramsey County, MN; Los Angeles County, CA; and Muskegon County, MI. This discussion focuses on the Ramsey County experience and touches on the experience in Muskegon County. The purpose in the Los Angeles County intervention was to improve the quality and timeliness of SSI applications, and hence should not be regarded as an early intervention demonstration.

Ramsey County piloted an intervention titled Families Achieving Success Today (FAST), comprising co-location of mental health services, health care services, and employment services, emulating the Individual Placement and Support (IPS) model.[18] In addition to the co-location of services, the intervention implemented case management and motivational interviewing. The target population for FAST was families receiving TANF benefits who were determined to not be making meaningful progress on TANF work requirements. Eligible individuals were randomly assigned to a treatment (FAST participation) or control group. To be eligible, the head of the family had to be between ages 22 and 59 and at least one member of the household had a disability.

Farrell et al. (2013) report outcomes that are characterized as exploratory for two reasons. First, the sample size for the FAST pilot was modest; 241 individuals in the treatment group and 148 in the control group. Second, only 63 percent of the treatment group received FAST services. Nevertheless, using TANF administrative data and wage record data, the evaluators found a slight reduction in TANF receipt (statistically significant for the first two quarters after enrollment) and a large statistically significant increase in average quarterly earnings. Unfortunately, there is no reported impact on SSI applications or receipt in the project report.

The Muskegon County program was intended to expedite the medical review used in TANF by relying on the SSI/SSDI Outreach, Access, and Recovery (SOAR) model[19] and to provide motivational interviewing and employment services to

[18] The IPS model is individual (client) driven, with services provided by a team usually within a mental health agency. When a client expresses interest in working, an employment specialist begins to meet with the client and to develop potential jobs. The client specifies whether the health issue should be disclosed. If employment is arranged, appropriate supports are continuously provided. The model has been shown in several randomized control trial studies to have statistically significant impacts on employment rates for individuals with severe mental illness. The IPS website (https://ipsworks.org/index.php/evidence-for-ips/, accessed May 12, 2021) enumerates 28 such studies of IPS impact. Among these, two recent studies (Baller et al. 2020; Hoffmann et al. 2014) document positive long-term employment impacts. Bond, Drake, and Pogue (2019) review studies that show that IPS improves employment outcomes for populations other than those with severe mental illness.

[19] SOAR is evaluated by Kauff et al. (2016).

individuals who were determined to be work ready with limitations. During the year-long pilot, only 68 individuals went through the SOAR-inspired medical review, and about one-third of the individuals determined to be work ready with limitations received employment services.

Supported Employment Demonstration (SED)

SED began in August 2016 and is scheduled to operate through August 2022 when final impact and benefit-cost analyses are due. The purpose of the demonstration is to see whether the IPS model can be effective in improving the labor market attachment of individuals with mental health conditions who applied for SSI or SSDI benefits and received initial denials or can reduce the number of their re-applications.[20]

The design of the demonstration is to randomly assign to one of two treatment groups or a control group 3,000 participants ages 18–49 recruited from the catchment areas of 30 community mental health centers. The demonstration's sites are geographically dispersed and vary by urban/rural characterization. The sites are in 21 states; 20 of the sites are urban and 10 are an urban/rural mix.

The treatment groups are designated as Full-Service and Basic-Service. Individuals assigned to the Full-Service treatment group receive the IPS employment services, services of a nurse care coordinator, systematic medication management, and assistance with cost sharing for medications and for behavioral health and work-related expenses for 36 months. Individuals in the Basic-Service treatment group receive the IPS employment services and cost-sharing assistance for behavioral health and work-related expenses for the same length of time, but do not receive the services of the nurse care coordinator or systematic medication management. Individuals assigned to the control group seek services as they normally would (or would not) in their community. Furthermore, at the time of randomization, each control group member received a comprehensive manual describing mental health and employment services in their local community, as well as state and national resources.

Enrollment for SED began in November 2017 and has been completed (Taylor et al. 2020). There have been no published impact analyses to date, but the final enrollment analysis report contains findings that could be relevant for early interventions. The enrollment procedure was complex, proceeding through the following steps: (1) SSA provided contact information for individuals in the appropriate catchment areas who applied for SSI or SSDI and were denied in the medical screening (*N*=73,512); (2) individuals determined to be ineligible for the demonstration were screened out (*N*=26,505); (3) a random sample of the remaining individuals was chosen to be contacted (*N*=21,003); (4) contact was made with about 65 percent of these individuals (*N*=13,375); (5) about 2,000 individuals were screened

[20] It should be noted that some of the individuals in the demonstration were denied SSDI because they had earnings that exceeded the SGA level. Thus, these individuals were attached to the labor market.

out for various reasons, leaving "potential enrollees" (*N*=11,307); and (6) approximately 26 percent of the potential enrollees enrolled in the SED demonstration (*N*=2,960).

Taylor et al. (2020) estimated an enrollment model using logistic regression. Assuming that the evaluation sample represents the target population, the estimates from this model give an indication of the characteristics of individuals who are most likely to take up a SED-like intervention. Among personal characteristics, men, individuals with higher educational achievement, and individuals with more limited work experience or earnings were more likely to enroll. Application denials due to evidence that the applicant could find alternative work in the national economy were strongly predictive of enrollment. Finally, two characteristics of the local labor market were associated with higher likelihoods of enrolling: higher unemployment rates and greater average wage growth.[21]

Retaining Employment and Talent after Injury/Illness Network (RETAIN)

Jointly with the US Department of Labor (DOL), SSA is administering the RETAIN demonstration. Its purpose is to test promising interventions that increase the labor force participation and potentially reduce the future need for disability benefits of individuals with recent serious injuries or illnesses. The project is largely based on a program operating in Washington State—Centers of Occupational Health & Education (COHE). The COHE program primarily addresses WC cases, whereas RETAIN extends the approaches used in Washington State to anyone in the labor force (not just WC cases) who has experienced an occupational or non-occupational injury or illness.

The RETAIN demonstration is being conducted in two phases. In the first phase, in summer 2019, eight states were awarded 18-month grants to launch pilot studies. In the second phase, five of the Phase 1 projects are being funded for full implementation and will be evaluated using an experimental design randomizing on individual participants in four states and a clustered random assignment design in the other state. Some states participating in Phase 1 of RETAIN target non-occupational injuries or illnesses only (KY, OH, WA), whereas others include work- and nonwork-related events (CA, CT, KS, MN, VT).

The RETAIN evaluation contractor is charged with implementing evaluations of the Phase 2 programs. To facilitate the design of the Phase 2 interventions, Anderson et al. (2020) documented the substantial state-level and county-level variation in SSDI and SSI application rates (applications as a percentage of estimated eligible individuals). From a national perspective, an interesting finding in this document is

[21] The COVID pandemic is having some effects on the demonstration. Services are continuing, although in most instances they are being provided online. Some participants have lost employment due to the pandemic, whereas others have gained employment (William Frey, pers. comm., October 2020).

the existence of a "belt" of states (mostly in the Southeast) that are estimated to have relatively high SSDI application rates. Early interventions might be expected to have the biggest impacts in these states.

As of now, there are no plans to report impacts/outcomes from the Phase 1 projects, which were intended to test the evaluability and program readiness of the smaller-scale projects. However, some evidence from COHE gives an indication of potential outcomes.[22] Funded by the Washington State Department of Labor and Industries, COHE provides early intervention and RTW services for individuals with work-related health conditions. There are six centers across the state, most of them housed in large medical systems. Injured workers effectively choose whether to use COHE services by receiving their care from a COHE-affiliated provider.

Health service coordinators (HSCs) are integral to the success of the COHE model.[23] HSCs work directly with injured workers, employers, health care providers, and other program participants to coordinate care and RTW activities for the injured workers. They monitor real-time data on all COHE cases and perform triage to identify cases that are likely to be long term or appear at risk of falling short of RTW goals. For cases needing assistance, they frequently contact injured workers, employers, providers, WC agency staff, and other stakeholders to facilitate the RTW process and to identify barriers to returning to work and resources for resolving them. The RTW activities they coordinate can include functional assessments, referrals to existing training and employment services, and setting appropriate RTW expectations. In the RETAIN demonstration, states' RTW coordinators fulfill this role with an increased emphasis on employment services. The coordination role is critical, as the program is based on the MacColl Chronic Care Model,[24] which asserts that a proactive system focused on keeping a person as healthy as possible will achieve greater success in that regard than will a reactive system.

An evaluation of the COHE pilot, which began in the early 2000s, showed promising results. COHE participants were less likely to be off work and on WC disability benefits one year after the initial WC claim, and combined medical and disability costs were reduced. The magnitude of these reductions was greater for back sprain cases, a common occupational injury, which likely influenced some of the RETAIN Phase 1 states to focus on musculoskeletal injuries. Franklin et al. (2015) report that at the eight-year mark, 2.5 percent of the COHE participants were receiving SSDI benefits, compared to 3.4 percent of the comparison group. This reduction is small in magnitude, but it is statistically significant.

---

[22] The description of COHE and its evaluation that follows above is excerpted from the funding opportunity announcement for the Phase 1 RETAIN grants (DOL/ODEP 2018).

[23] According to a reviewer of this chapter, COHE staff also identify the existence of a centralized data base that is easily accessible by stakeholders as a major driver of that program's success.

[24] For a description of the model, see https://maccollcenter.org/resources/chronic-care-model (accessed May 12, 2014).

Promoting Work through Early Interventions Project (PWEIP)

PWEIP is a funding collaboration between SSA and ACF. ACF has initiated two projects targeting low-income individuals with little or no work history, with current or foreseeable disabilities, who have not applied for SSI. The two projects are Building Evidence on Employment Strategies for Low-Income Families (BEES) and Next Generation of Enhanced Employment Strategies (NextGen).

BEES

The BEES project will test various strategies designed to improve the labor market outcomes of individuals with low incomes and barriers to work. The primary employment barriers that the BEES project targets:

- Substance and opioid use disorder;
- Criminal justice involvement; and
- Mental health and disability issues.

According to ACF's Office of Planning, Research, and Evaluation (HHS/ACF/OPRE 2020), BEES has identified eight sites for further study or evaluation: Addiction Recovery Care (KY), Breaking Barriers San Diego (CA), Central City Concern (OR), IPS within Federal Qualified Health Centers (IL, NH), The Journey (OH), Substance Use Disorder Sites (multiple states), Two-Generation Residential Mobility Demonstration (IL), and WorkAdvance (multiple states).

Breaking Barriers San Diego appears to be a continuation of data collection from a DOL Workforce Innovation Fund (WIF) project that was evaluated using random assignment, as documented by Freedman, Elkin, and Millenky (2019). That project embedded the IPS model in a workforce setting as opposed to the mental health setting for which the IPS model was developed. Besides the effort in San Diego, another BEES site started random assignment in 2019 but has curtailed enrollment due to COVID.[25]

The BEES project will continue to work with states to identify effective interventions targeted on families with low incomes (HHS/ACF/OPRE 2020). When interventions have been identified, ACF and its contractors will implement the most rigorous evaluation approaches, focusing on random assignment where possible.

NextGen

The second project is planning to evaluate interventions at nine sites. The intended interventions will target individuals with current or foreseeable disabilities who have limited work histories and are at risk of applying for SSI. Because it has been recognized that employer involvement is a key element in successful job training,

[25] K. Martinson, email with the author, December 2020.

NextGen will attempt to include interventions that involve employers or are market-oriented approaches. To date, five interventions have been selected to participate in NextGen: Bridges from School to Work (eight urban areas); Families Achieving Success Today (Ramsey County, MN); IPS for Individuals with Justice Involvement (pending at selected mental health centers); Work Success (Utah Department of Workforce Services); and Wellness, Comprehensive Assessment, Rehabilitation, and Employment (New York, NY) (HHS/ACF/OPRE 2020). Findings on the effectiveness of these interventions are likely to be released beginning in 2023.

INTERNATIONAL EXPERIENCE

Several European countries have enacted policies aimed at reducing expenditures on disability benefits.[26] The following section provides brief descriptions of the policies and, as available, their outcomes. The lessons learned from the experiences could be instructive in considering early intervention policies and practices in the United States.

The Netherlands

The Dutch government responded to rapidly growing rolls and costs of the country's disability insurance system in the 1980s and 1990s by having employers bear some of the costs borne by the system when their workers made disability claims. Starting in 1994, the government required all employers to finance the first six weeks of their employees' sickness benefits.

Two years later, the government lengthened the time to one full year. These reforms continued in 2002 with the introduction of the "Gatekeeper Protocol," which required the employer, worker, and a consulting physician to jointly draft a return-to-work plan within eight weeks of a disability claim and appoint a case manager to coordinate this process. In 2004, mandatory employer-paid sickness benefits were extended from one year to two years, as was the mandatory waiting period for access to public disability benefits. Thus, employers retained full financial responsibility for their employees' sickness benefits for two full years.

After these two changes were implemented, along with the full phase-in of experience-rated disability insurance premiums,[27] the inflow of participants into the Dutch disability program fell by 40 percent from 2002 to 2004 and by another 50 percent from 2004 to 2006. Whether the changes caused the drops is unknown. A likely downside to the Dutch reforms is the inadvertent reduced likelihood of hiring

---

[26] Main sources for this section are Burkhauser et al. (2014) and SSA (2018b).

[27] "Experience rating" an insurance program means that employers with fewer workers entering the program would pay a lower premium, and those with more workers entering the program would pay a higher premium.

individuals with a disability because of the potential employer responsibility for benefits (Hullegie and Koning 2015).

Sweden

In 2003, the Swedish government merged its sickness benefits and disability systems and began a series of changes to standardize and enforce the administration of this now joint system. Most notable among the changes was the centralization of screening processes. Until then, many doctors and regional disability gatekeepers had focused the support that they provided to injured individuals on providing income support, rather than work retraining. By centralizing the process and developing standardized protocols for granting cash benefits, policymakers were better able to regulate the gatekeepers and enforce a strategy of promoting participation in work before offering cash benefits. In addition to standardizing the screening process, employers were required to meet with disability administrators to create a rehabilitation plan. And administrative gatekeepers were given the power to demand that employers provide certification about the types of accommodations they made for the worker.

In 2008, the Swedish government further reformed the sickness benefits program, which reduced the flow of applicants to the long-term disability system. Frequent checkpoints were established that included work capacity assessments, and cash benefits were reduced for those who did not return to work. Earlier checkpoints provided rehabilitation, counseling, and assessment much closer to the onset of an impairment, when return to work was more likely.

The reforms increased the return to work of new sickness program entrants and reduced their time on the program. In contrast, few of those already on the sickness program when the new reforms were initiated ever returned to work. When their sickness benefits ended, they simply moved onto other social assistance programs. These findings provide empirical evidence that early intervention matters.

Great Britain

A period of rapid growth in disability receipt rates came to an abrupt end in 1995 with a set of major reforms that ended the Invalidity Benefit (IVB) program and replaced it with the Incapacity Benefit (IB) program for all new beneficiaries. IB was less generous than IVB, and medical screening was now carried out by government doctors working for the relevant agency rather than by family doctors. The bar was also set higher, moving to an assessment of the claimant's capacity to carry out *any* work rather than work in their usual occupation.

Great Britain piloted a work-first reform called Pathways to Work in 2003 and rolled it out nationally in 2005. It made movement onto the disability benefits program conditional on attendance at work-focused interviews, introduced a "back to work"

bonus payment, and provided additional in-work supports for those returning to employment.

In 2008, the Employment and Support Allowance (ESA) program replaced IB. A new tougher Work Capability Assessment (WCA) with few exemptions was a feature of the ESA, which is an insurance-based benefit for those with sufficient work history. The WCA was also required of individuals without sufficient work history who received the pre-existing means-tested social assistance benefit. The WCA triages ESA applicants into groups identified as Fit for Work, Work-Related Activity Group, and Support Group. Members of the first group do not receive disability benefits. Members of the last group are individuals with severely limiting disabilities, and they receive a full disability benefit. Members of the Work-Related Activity Group are assessed as having limited capability for work; they receive a time-limited benefit that is approximately three-fourths of the benefit received by the Support Group members. Even though the ESA and tougher WCA were substantial reforms, they have not achieved the government's benefit receipt reduction goals nor had success in helping people with disabilities stay in or enter the labor force (Inanc and Mann 2019).

EARLY INTERVENTION STRATEGIES SUGGESTED IN THE LITERATURE

To add to the modest amount of evidence available from demonstration initiatives and from the policies pursued in other countries, this section of the chapter presents and critiques a number of early intervention strategies that have been suggested in papers but have not been implemented in a demonstration or policy.

Reforms Suggested Post–Great Recession (2010–2013)

Autor and Duggan (2010) proposed requiring employers to provide disability insurance, in the same way that employers are mandated to fund the UI and WC programs. In particular, they argue for universal, experience-rated PDI coverage with minimum standardized benefits. The benefits would include VR services, workplace accommodations, and partial wage replacement. Autor and Duggan provide a detailed analysis of such a proposal that includes the suggestion that the cost to employers would not be prohibitive. However, mandating PDI is well beyond the purview of SSA. Accomplishing this reform would require federal legislation or legislation in all states, neither of which is likely to happen. Also, the findings of Stepner (2019) offer a concern. They indicate the take-up of short-term disability insurance leads to increased usage of long-term disability insurance.

Burkhauser and Daly (2011) propose a system of experience-rating the SSDI portion of the FICA payroll tax for employers, as well as devolving SSI back to the states. They argue these differential rates would encourage employers to do more to retain workers at risk of becoming disabled, including work accommodations, rehabilitation services, and return-to-work efforts. Like the Autor and Duggan (2010)

proposal, their suggested reforms are structural and would require substantial legislative action.

Liebman and Smalligan (2013) propose three less ambitious initiatives that they suggest could be implemented by SSA as demonstration programs. The first is to offer a package of benefits to SSDI applicants who have been determined through a screening process to be work ready if provided supports, in exchange for their suspending their application. The benefits would include targeted vocational and health interventions, an Earned Income Tax Credit–like wage subsidy, and potentially, a few months of an emergency cash diversion grant. The cost-effectiveness of this idea would depend on the effectiveness of the screening in targeting supports to applicants who would otherwise receive SSDI. The proposal is silent, however, on how this screening would be accomplished. Furthermore, the offer of a cash diversion grant might induce individuals to apply who otherwise would not do so.

The second initiative would be to allow states to reorganize the federal share of existing social program funding streams (VR, TANF, community mental health, Medicaid) to target populations that are likely to end up receiving a lifetime of SSI or SSDI benefits in the absence of assistance. Further, states could receive incentive funding if they demonstrated success at reducing participation in SSDI and SSI. This proposal attempts to fund a demonstration out of existing program funding, which would presumably be replaced with incentive funding from SSA if the state's interventions are shown to reduce SSDI or SSI costs. It is not clear why the state-level administrators of the programs from which funding is taken would benefit from this demonstration, and why they would risk losing funding if the demonstration is not successful.

The third would be to provide a tax credit against employers' disability insurance payroll tax for firms that can reduce the disability incidence of their employees by at least 20 percent. The authors themselves raise the possibility that financial incentives to employers might result in discriminatory hiring against workers with disabilities. They suggest that it would be worthwhile to conduct a demonstration in order to learn about such a practice.

McCrery-Pomeroy SSDI Solutions Initiative (2016)

An activity of the Committee for a Responsible Federal Budget, the McCrery-Pomeroy SSDI Solutions Initiative commissioned several studies that offered reform ideas for various aspects of the SSDI program. Among these were four studies that offered early intervention strategies, described briefly here.

Stapleton, Ben-Shalom, and Mann (2016) propose a new institution that they have named the Employment/Eligibility Service (EES) that would integrate workforce supports with SSDI eligibility determination. They envision the EES as the organization that an individual who experiences an employment-threatening injury or illness would contact. For individuals who have reached insured eligibility, an adjudicator would assess the likelihood that the individual could return to work with

available supports. If not likely, the individual would be awarded SSDI (contingent on SSA review). If likely, then supports would be offered that might include development of a work plan, health care services, rehabilitation, accommodations, assistive technologies, transportation assistance, personal assistance, trial or gradual return to work, employer incentives, and cash assistance.

Christian, Wickizer, and Burton (2016) also propose a new institution that they name the Health & Work Service (HWS) that would respond quickly when individuals are having difficulty coping with a work-threatening impairment. Like COHE and RETAIN, the HWS would employ a coordinator who will:

> facilitate communications and problem solving among the key parties; identify issues that require attention; refer outside for special expertise or outside resources; coordinate care and services as needed; and provide positive support for the affected individuals, guiding them toward functional restoration so they can stay at or return to work. (94)

The HWS would intervene with evidence-based services within 12 weeks of the medical episode.

Kerksick, Riemer, and Williams (2016) propose piloting a transitional jobs initiative for potential SSDI applicants as well as for beneficiaries. The initiative would be administered by an intermediary that would become the "employer of record" for individuals who are placed in a transitional job. Such jobs would be subsidized (up to $10 per hour), wage-paying, full- or part-time jobs, typically in the private sector. The jobs would be available to individuals who have been out of work four weeks or longer and would last up to six months or 1,040 hours. While holding the transitional job, individuals would interact with a job counselor, who would help place the individuals at unsubsidized jobs. In addition to the transitional jobs initiative, these authors propose an expanded Earned Income Tax Credit and regular access to a work incentives counselor.

Manchester (2019) suggests that states should carefully analyze the medical records of SSDI beneficiaries and SSI recipients to see where early intervention initiatives might be targeted. She shows that states vary considerably in their share of those with mental disorders or substance use disorders. Accordingly, early interventions aimed at these disorders are likely to be most efficacious in states with high incidences.

Ekman (2016) formally critiques the first three papers. She suggests that the new institutions being suggested by Stapleton, Christian, and their colleagues will not be affordable and will be radical changes that will not be accepted by disability advocates or existing program staff. She furthermore argues that the cost savings assumptions in the Stapleton et al. paper are not achievable. Her critique of the Christian proposal is that it is mainly guided by WC program experiences and that it doesn't adequately address medical conditions that develop over time. Her suggestion for the Kerksick

transitional jobs initiative is that it be bolstered by a refundable credit for impairment-related work expenses.

Over and above the critiques offered by Ekman, a number of elements suggested by these papers seem to raise questions as to their viability as early interventions. The EES proposed by Stapleton, Ben-Shalom, and Mann seems to be addressed to SSDI, and it is not clear how SSI applicants would be handled. Furthermore, the underlying assumption behind the proposal is that a new institution that integrates workforce supports with SSDI eligibility determination would be more effective at getting SAW/RTW services delivered than the current "system" is. This seems like a strong assumption that is essentially untested and relies on integrating two functions that to date have not been housed together.

The RETAIN demonstration is essentially testing the idea of a coordinator of services advocated by Christian and colleagues, although RETAIN is not going as far as forming an entirely new institution. As Ekman alludes to, a question that needs to be addressed for the proposed new institution as well as for RETAIN is how individuals who experience a medical event that is not work related or that develops over time would be brought into the system. The transition jobs notion in the Kerksick et al. paper is perhaps novel for assisting individuals with disabilities, but it is silent on the source of the overall funding for this proposal and how the proposal would be administered. It relies on identifying intermediary organizations that would develop the jobs and would be the employers of record. The intermediary organizations would need to monitor and enforce provisions that prohibit employers from displacing existing staff or replacing workers in labor disputes. Furthermore, given that the transitional jobs will last no more than six months, it is not clear whether a large number of such jobs could be developed in the private sector, especially if workplace accommodations are necessary.

Finally, the Manchester paper identifies a useful source of information to guide policymakers in terms of geographic locations for early interventions, but it does not provide suggestions for what early interventions might be implemented.

WHAT HAVE WE LEARNED?

Ekman (2016) makes the following statement: "There is neither completed research nor an evidence base upon which to enact nationwide early intervention or work support programs" (134). This chapter's review suggests that this statement still holds, although the rigorous demonstrations in progress and some evidence from completed demonstrations and from the international arena have yielded lessons. Furthermore, a number of analysts have proffered thoughtful ideas about early interventions that merit further consideration and testing.

To date, only two experimental evaluations have reported the impact of early intervention strategies on SSDI or SSI applications or benefits. The DMIE evaluation showed a statistically significant reduction in applications and benefits in one of its sites (TX), and the Breaking Barriers San Diego site, funded by a WIF grant, had no

detected impact. Though it was not experimental, the COHE evaluation reported a reduction in SSDI beneficiaries over an eight-year period. As noted above, the reduction was small in magnitude and the program covered only WC beneficiaries. Despite this paucity of evidence, there are some findings from the descriptive or evaluation analyses conducted within the demonstrations to date or from the international experiences that point to promising principles. These findings are enumerated in the following section.

Elements of Early Interventions

Coordinator/Case Manager[28]

Evidence suggests that a key element for an early intervention initiative is the assignment of the coordination of activities to an individual. The two-phased RETAIN demonstration is following up on successful SAW/RTW initiatives conducted in Washington State. In particular, COHEs have operated and have been evaluated there. Analyses show that the COHE model statistically significantly reduced SSDI beneficiaries among individuals injured on the job. Integral to the COHE model is a health services coordinator who coordinates care for the injured individual. The coordinator acts as a case manager and interacts with the individuals with injuries or disabilities, employers, health care providers, and other parties as appropriate. RETAIN has required sites to engage an RTW coordinator to fulfill this role. The Dutch Gatekeeper Protocol, in which a case manager is required, also suggests that having a coordinator matters. Although its results will not be known for several years, the RETAIN demonstration will provide SSA and DOL with evidence of the effectiveness of the projects designed by the participating states. These projects are loosely based on the COHE model principles. A consistent element across RETAIN's state projects is the active involvement of a return-to-work coordinator to facilitate continued employment.

Timely Intervention

Much of the SAW/RTW literature emphasizes intervening after a work-threatening medical event as quickly as possible because the likelihood of returning to work decreases meaningfully with time. The RETAIN demonstration requires early communication to all stakeholders to return the worker to the workplace as soon as possible. The Dutch Gatekeeper Protocol requires the employer, worker, and a consulting physician to jointly draft a return-to-work plan within eight weeks of a disability claim. In Sweden, when reforms set up early checkpoints in its sickness benefits program, new beneficiaries had much higher return-to-work rates than the

---

[28] Chapter 8 in this volume provides additional discussion about the role of coordinators/case managers.

beneficiaries who were already in the program when the reforms were introduced, a finding that reinforces the importance of early intervention.

Individual Placement and Support (IPS)

The IPS model, which was developed for individuals with mental health conditions and which has been found to be quite successful at improving labor market outcomes, has been implemented in two demonstrations discussed in this chapter. One of the demonstrations (SED) is evaluating the efficacy of IPS in a number of sites, but that demonstration is in process and there have not been any outcomes observed yet. In the TSDTP, Ramsey County (MN) implemented FAST, a version of IPS, and the published results from that demonstration indicate employment and earnings impacts, but no evidence on disability benefit receipt (Farrell et al. 2013). At least one of the sites identified for further study in PWEIP's BEES demonstration and two of the sites identified in its NextGen project are focused on IPS as the intervention.

It should be noted that the IPS model was shown to improve employment outcomes in the MHTS (Frey et al. 2011), but the treatment did not reduce SSDI benefits. The IPS intervention in the WIF-funded Breaking Barriers San Diego site found no statistically significant differences in any of the main outcomes—not in employment or earnings, nor in the share of participants receiving SSI or SSDI (Freedman et al. 2019).

Employer Responsibility

Several European countries that were experiencing burgeoning disability benefit rolls and costs enacted reforms that seemed to reverse the trends. In the Netherlands and in Sweden, reforms placed more responsibility for financing sickness or disability benefits on the shoulders of employers, giving them a greater incentive to assist workers in staying at or returning to work. The Dutch require employers to maintain the payment of virtually 100 percent of an injured/disabled worker's earnings for up to two years, and these payments are experience rated. Sweden passed similar reforms. It is notable that the NextGen project is attempting to find early interventions that involve employers.

Moving to a system in which employers bear more responsibility, such as has been done in the Dutch and Swedish cases and as is being suggested with experience rating SSDI in the United States, has issues to confront. As noted, the Dutch experience resulted in reduced hiring of individuals with disabilities. Furthermore, whereas experience rating WC addresses work-related disabilities or illnesses, it is not clear that employers should shoulder the costs of disabilities resulting from nonwork-related events.

Using Data to Identify Likely Effectiveness of Early Interventions

Analyses of data from demonstrations or administrative sources may help to identify effective early interventions. Taylor et al. (2020) provide descriptive analyses of the characteristics of individuals who chose to enroll in the SED after being given an overview of the intervention. Assuming that this decision is analogous to the decision to participate in an early intervention for individuals with mental health conditions, it suggests that the following personal characteristics increase the likelihood of participation: men, individuals with higher educational achievement, and individuals with more limited work experience or earnings. Application denials due to evidence that the applicant could find alternative work in the national economy were strongly predictive of enrollment. Two characteristics of the local labor market were associated with higher likelihoods of enrolling: higher unemployment rates and greater average wage growth.

If SSA were to implement an early intervention, it would want that initiative to be as target efficient as possible. That is, the agency would not want to spend resources on individuals who were unlikely to apply for benefits in any case. For example, evidence suggests that targeting early intervention strategies *by state* could make sense. Anderson et al. (2020) document the existence of a "belt" of states (mostly in the Southeast) that is estimated to have relatively high SSDI application rates. This finding suggests that to the extent that regional variation arises in an early intervention strategy, then these states might represent where biggest impacts could be expected. Similarly, Manchester (2019) used medical records data to show variation by state in the share of SSDI beneficiaries and SSI recipients with mental disorders or substance use disorders.

Stapleton et al. (2015) document characteristics of the target population for which evidence of early intervention effectiveness has been shown. In a review of studies, they summarize the characteristics of individuals for whom the studies indicated early intervention(s) had positive employment outcomes. These characteristics include individuals with musculoskeletal conditions, especially lower back pain; individuals with mental health conditions; individuals with chronic conditions for which adherence to treatment is critical; and individuals who remain attached to an employer. They also cite a Dutch study in which male workers ages 40–58 had more positive outcomes than did younger men.

Caveats

To date, the evidence base is extremely thin. The ideal situation in implementing an early intervention would be having evidence from rigorous evaluations of multiple, externally valid demonstrations. The RETAIN, SED, and PWEIP demonstrations are arguably important strides in the right direction of filling our knowledge gaps about the effectiveness of potential early interventions. Their impacts on SSDI or SSI

application rates will not be identified for several years, however. So we need to consider the lessons learned to date in the following light:

- The experiences in European countries might not translate to the United States because of different employment relationships, institutions, demography, politics, economy, and other international variations.
- DMIE was conducted in just a few states, with variation in the treatments and at a time when the economy was strong.
- The findings about the IPS model presented in this chapter have limitations. The WIF-funded Breaking Barriers San Diego program, which found insignificant impacts for labor market and disability program benefits, was conducted at a time (2016–2019) and in a single metropolitan area with a strong economy. Furthermore, the IPS treatments that were tested in that program and in Ramsey County (MN) in TSDTP were situated in a workplace setting rather than a mental health setting, where studies have found more robust findings.

FUTURE DIRECTIONS: A PROPOSED DEMONSTRATION

In considering what has been learned to date, there seems to be an area in which SSA might consider conducting an early intervention demonstration. The idea is to conduct a demonstration that targets individuals age 50 or older who are denied benefits. In other words, this would be a SED-like demonstration for older applicants. Such applicants are going to have two barriers to overcome if they choose to search for re-employment: age and disability. Furthermore, it could be the case that these individuals have not actively searched for employment for many years.

The treatment in such a demonstration would be specific workforce development strategies for older workers. These would include job search assistance as well as job development. To implement such a demonstration, SSA could announce a funding opportunity for workforce agencies that serve older adults for specific projects that would serve individuals with disabilities. Interventions and supports might include case management, career counseling, job search assistance, training, or job development that might even include transition jobs as proposed by Kerksick, Riemer, and Williams (2016). The types of agencies that might be interested in developing disability-targeted approaches for seniors include agencies administering the DOL-funded Senior Community Service Employment Program, the AARP Foundation's Back to Work 50+ program, or VR agencies.

Agencies chosen by SSA would receive contact information for individuals older than age 49 whose applications for SSI or SSDI have been denied. Those individuals would be contacted and given the choice to participate in the demonstration. The individuals who volunteered to participate would receive information about programs in the area serving seniors and would be randomly assigned to treatment or control groups.

Developing employment opportunities for older individuals with disabilities is challenging. However, a study of the Senior Community Service Employment Program, which served individuals older than age 55, found an employment rate of 52 percent for the total population served and 37 percent for individuals with disabilities (Kogan et al. 2012). Unfortunately, that study did not include any sort of control or comparison group, so achieving an employment rate of 37 percent for individuals with disabilities age 55+ could have been quite a success. At a minimum, setting up a rigorous evaluation for this demonstration to learn what workforce development strategies can or cannot work for seniors would be extremely valuable. It would also be important to track the disability re-application rates to determine the efficacy of this type of early intervention.

CONCLUSIONS

In the United States, there seem to be two options on the menu of early interventions. The RETAIN approach attempts to standardize and coordinate the SAW/RTW activities at a regional level, whereas the IPS model places responsibility on the individual. These options are not in opposition to nor mutually exclusive of each other. RETAIN serves occupational and non-occupational medical events that involve physical or mental disabilities. The IPS model has been shown to be effective in improving labor market outcomes for individuals with mental disabilities. Two large-scale demonstrations of these approaches are in process—RETAIN and SED—with impact results available in a few years. Under the PWEIP umbrella, the BEES and NextGen projects are getting underway, and they will be testing employment strategies such as IPS aimed at overcoming barriers to quality jobs that some individuals may face, such as mental health issues, substance use disorders, or other barriers.

Some European countries have had success in stemming the inflow of disability benefit applicants by requiring employers to bear the costs of sickness/disability benefits. Presumably, this incents employers to assist employees in staying at work after major medical events. Furthermore, these countries as well as some US reform ideas suggest that employer costs be experience-rated. However, given the labor market institutions and political sway of the employer community in the United States, these sorts of cost shifting seem unlikely here. Furthermore, shifting costs to employers could exacerbate discrimination against workers with disabilities; one European study found this result.

Similarly, the reform ideas involving mandated (short- or long-term) disability insurance schemes seem unlikely to be accepted in the United States. Employer mandates that might arguably increase costs are unlikely to find advocates and are usually political nonstarters, although, of course, political winds are subject to change. On the other hand, it should be noted that there are five states with mandatory temporary disability insurance programs, and it is probably the case that promulgating mandatory UI or WC could have seemed infeasible when they were initiated.

While we wait for impact results from the large-scale early intervention demonstrations that are ongoing, there are some incremental lessons to heed. Early interventions should:

- Take place as soon as possible after a work-threatening injury or illness occurs;
- Be case managed/coordinated;
- Involve health care professionals who have been trained in and accept staying at work or returning to work as a desirable treatment outcome; and
- Target individuals/regions with characteristics that data suggest are likely to succeed.

Waiting will take patience that will be rewarded with solid evidence about the effectiveness of the strategies being demonstrated.

Chapter 5

Comment

Jeffrey B. Liebman
Harvard University

Let me begin by complimenting Kevin Hollenbeck for writing an excellent chapter ("Demonstration Evidence of Early Intervention Policies and Practices"). To synthesize such a wide range of past and ongoing evaluations in such an insightful manner is quite impressive.

Let me also observe how terrific it is to see the Social Security Administration (SSA) conducting so many innovative demonstrations and partnering with other federal agencies on many of them. As Hollenbeck notes, by the time SSA encounters a Supplemental Security Income (SSI) or Social Security Disability Insurance (SSDI) applicant, the ideal time to intervene may well have passed. So it is important for SSA to be working with other agencies that might encounter future applicants further upstream and who have deep expertise in the health and employment aspects of interventions. These initiatives show our federal government at its best as a learning and continuously improving organization, one that is capable of breaking down agency silos to provide better services.

That said, I have two concerns about SSA's early intervention learning agenda.

UNDERPOWERED EXPERIMENTS

Hollenbeck notes that sample sizes for some of the experiments have been "modest." When experiments are too small, it is hard to learn anything conclusive from them. This is especially true if one does the correct adjustments of confidence intervals for the fact that there are multiple outcomes being measured, with results often presented separately for different sites. Sometimes one needs to make the tough call and not go forward with a 5- or 10-year experiment, no matter how innovative, if at the end of the day budget constraints or sample recruitment challenges mean that the results are almost certainly going to be inconclusive. Of course, the best solution to this problem is to provide SSA with the resources necessary to do experiments with adequate statistical power.

Relatedly, a challenge in developing successful early intervention programs is that the population of people with work-limiting disabilities is quite heterogeneous. Intervention strategies will often need to vary by health impairment and occupation. Given sample size limitations, we are likely to make more progress on SSA's learning agenda if we develop focused interventions for some of the most common health impairments and job types, rather than developing broad strategies and then trying to estimate subgroup impacts. For this reason, Hollenbeck's suggestion that SSA develop an initiative targeted specifically for workers in their 50s is sensible. In addition to representing a large portion of disability insurance applicants, such workers have

enough potential remaining years of employment to produce a stream of benefits that exceeds the upfront intervention costs.

UNCLEAR MOTIVATION

We need to be clearer about what is motivating us to do early intervention. One view is that we are trying to increase economic output, and therefore our nation's standard of living, by putting more people to work. Another view is that we are trying to reduce government spending by diverting people from claiming benefits. A third view is that we are trying to improve the well-being of people who are struggling with both health impairments and labor market challenges by helping them get back on their feet.

If our primary motivation is either of the first two, we are likely destined to fail. The labor market prospects of people on the margin between receiving and not receiving SSDI and SSI benefits are not all that great, even in the best of circumstances; often the number of extra years in the labor force that can be expected even if someone returns to work is not all that high; the interventions are expensive; and if one does the benefit-cost analysis properly and subtracts the workers' disutility of effort from the output gains, it is very unlikely we will design an intervention with social benefits that exceed costs.

The same is true if our motivation is government finances. Early intervention programs typically serve many people per person diverted from benefit receipt, so it is difficult for an intervention to fully pay for itself. Moreover, given that the target population consists largely of workers with low incomes struggling with health impairments and other challenges—people who deserve a high social welfare weight—we would have to believe the "leaky-bucket" of our income transfer system is very "leaky" in order to think we are doing good when we reduce benefit spending.

Thus, I would argue that the main reason we should be designing, implementing, and evaluating early intervention programs is to improve the well-being of those to whom we are providing services. This perspective has at least four important implications.

First, our primary outcome measures in these studies should be measures of well-being—pain levels, depression levels, substance use levels, divorce and domestic violence levels, happiness, and longevity, among others. Employment and benefit receipt may in some cases be useful proxy measures, but they should not be the main or only focus. In addition to being conceptually right, taking this approach to measuring a broader set of outcomes also makes it much more likely that we will find benefits of an intervention that are substantial enough to exceed costs.

Second, I think a lot of us, me included, have a presumption that when we help someone get back to work we are indeed doing something good for them. And conversely, that in telling someone we will give them lifetime benefits in exchange for never working again we may be in many cases consigning people to misery. But we really have not done the research necessary to know whether this is right on average,

much less for which subpopulations this is correct. Someone should fund a major study using the Maestas, Mullen, and Strand (2013) disability examiner instrumental variable to compare well-being impacts of receiving versus not receiving SSI and SSDI. Because the study would need to collect most of the outcome data directly from participants rather than by using administrative records, it would probably cost $10 to $20 million to do this right.

Third, in all of our early intervention studies (and indeed in most social experiments) we should have an extra experimental arm where we simply give people extra cash for a few years equal to the budgeted per capita amount it costs to deliver the intervention. In determining whether our interventions are effective, we should be held to the standard that not only do our interventions work, but that they work better than giving people the same amount of cash.

Fourth, we should test a guaranteed income approach to disability benefits. I am not a fan of giving a guaranteed income to everyone in the United States. The amount of extra taxes it would take to fund such a program is prohibitive. But a guaranteed income for low-earners with health impairments who are struggling in the labor market is much more appealing. We should take one state and provide SSI and SSDI benefits to a targeted set of low-earners meeting the standard qualifications for the programs—but free of any limits on subsequent employment and with ongoing health insurance guaranteed, as well. Doing so would almost certainly increase benefit applications. If it improved well-being and caused applications in the target population to double, I personally would think we had done a good thing. If it caused applications to rise 10-fold or if it led to more people in despair because of lack of purpose, I would think it was a disaster. Only an evaluation can help us determine which is more likely.[29]

[29] Such an evaluation cannot be done under SSA's current demonstration authority. It would require new authority and funding from Congress.

Chapter 5

Comment

Jennifer Sheehy
US Department of Labor[30]

The chapter authored by Hollenbeck ("Demonstration Evidence of Early Intervention Policies and Practices") is an excellent summary of the state of the science on early intervention. He cites evidence for strategies to improve stay-at-work or return-to-work outcomes with early interventions, but concludes that there is no one-size-fits-all solution. Hollenbeck describes five US-based early intervention demonstration programs that have been administered or supported by the Social Security Administration (SSA), as well as relevant international efforts. Early intervention programs can take many forms, and the programs described differed in terms of services offered, intervention timing, and participant characteristics.

Summarizing lessons learned from past demonstrations, early interventions should:

- take place as soon as possible after a work-threatening injury or illness occurs,
- include case management and coordination,
- involve health care professionals who have been trained in and accept staying at work or returning to work as a desirable treatment outcome, and
- serve individuals/regions with characteristics that data suggest are likely to succeed.

Because the exiting evidence base is thin (Ben-Shalom et al. 2017), we are all looking forward to findings of ongoing early intervention demonstrations, especially Retaining Employment and Talent after Injury/Illness Network (RETAIN), the new demonstration conducted by the Office of Disability Employment Policy in collaboration with the Department of Labor's Employment and Training Administration and being evaluated and partially funded by SSA.

RETAIN provides early coordination of health care and employment services through an integrated network of partners. Its goals include improving the employment outcomes of newly injured or ill workers and reducing the need for Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). RETAIN will develop evidence on the effectiveness of early intervention stay-at-work and return-to-work efforts. The programs are modeled after Washington State's Centers of Occupational Health & Education (COHE), but seek a broader target population that

[30] The views expressed in this chapter are those of the author and do not necessarily represent the views of the Department of Labor or the US federal government.

includes those with non-occupational injuries/illnesses and provide a more expansive set of services to injured and ill workers.

Hollenbeck describes RETAIN as primarily operating among health care providers and employers and, though these stakeholders play critical roles in RETAIN, many insurers and social programs listed by Hollenbeck also play key roles. Each RETAIN grantee has a workforce partner, which enables the coordination of health care and employment services; and many state RETAIN programs serve individuals who are receiving insurance benefits through workers' compensation, private disability insurance, and others.

The theory of change for RETAIN is based on evidence that the probability of returning to work after missing 12 weeks of work drops dramatically (IAIABC 2016). RETAIN is targeting individuals with a connection to the workforce, with the goal of providing services within 12 weeks of work disability onset. Further, RETAIN participants may not have applied for or be receiving SSDI or SSI benefits. This means effective services for people out of work for extended periods may differ from services offered as part of RETAIN. Though RETAIN was initially focused on workers with musculoskeletal conditions, most programs have expanded to serve workers with any condition that inhibits their work.

That the probability is low of returning to work after missing 12 weeks of work suggests the early stages after work disability onset shape the trajectory of the worker's outcome. Health care professionals play critical roles in the early stages, though they are typically not trained in occupational health best practices and may not be thinking of work as a positive health outcome (Denne, Kettner, and Ben-Shalom 2015). To address this, RETAIN programs train health care providers in occupational health best practices and incentivize the providers to adopt those best practices.

The systems that serve individuals at risk of dropping out of the labor force and/or applying for SSDI/SSI are fragmented and typically do not coordinate. Workers may receive wage-replacement benefits from insurers, services from health care or rehabilitation providers that treat their health condition; employers may provide job accommodations or stay-at-work/return-to-work services; and workers may seek education and training to perform a new job. These interactions influence a worker's ability to stay in their current job or return to the workforce, but each stakeholder has its own goals and incentives which are not always aligned with the goal of keeping individuals in the workforce (Epstein et al. 2020). RETAIN is seeking to align stakeholder incentives around the goal of helping injured and ill workers recover and return to the workforce.

The target population is diverse and challenging to reach (Nichols et al. 2020), so RETAIN is engaging employers and establishing policies to integrate key networks to help workers stay at or return to the workforce after injury or illness. Other potential approaches include integrating stay-at-work/return-to-work services into paid family and medical leave policies and providing targeted stay-at-work/return-to-work information to workers, employers, and medical professionals. RETAIN also includes

longitudinal survey and administrative data analysis to learn more about this diverse population, but more analysis is needed to understand how best to serve people struggling to get back to the workforce after experiencing an injury or illness. By providing holistic care that focuses on work as a positive health outcome, effective early intervention services may help people with the ability and desire to continue working.

Chapter 6

Youth Transition

David Wittenburg and Gina Livermore
Mathematica

The transition to adulthood for youth (ages 14–25) receiving Supplemental Security Income (SSI) is a subject of strong policy interest. These youth face potential barriers related to their health and limited resources that can affect their access to opportunities. Additionally, the SSI eligibility rules change at age 18 from a child-based definition to an adult-based definition. The planning for this transition, especially for the potential loss of benefits, is a major concern for families. On average, almost half of the income for families of children with disabilities comes from SSI (Davies, Rupp, and Wittenburg 2009), which means the potential loss of SSI can have an impact on family resources.[1] The combination of health, resource, and program eligibility issues can create challenges for youth in pursuing activities that can further their development. For example, if a youth engages in substantial work, he or she might no longer qualify for benefits.

Two issues motivate policy interest in providing support during the transition from youth to adulthood. First, the evidence indicating that youth receiving SSI face difficulties in their adult years with employment, education, and independent living outcomes drive interest in improving transition supports (Deshpande 2016a; Hemmeter, Mann, and Wittenburg 2017; Wittenburg 2011). Second, the SSI child caseload has grown relative to other programs that provide income support to low-income families. However, the growth in SSI participation has been consistent with other programs that provide in-kind support, such as Medicaid. Nonetheless, the overall changes in the delivery of SSI benefits have raised questions about ways to provide income supports that have been the subject of policy interest (Boat, Buka, and Perrin 2015; Duggan, Kearney, and Rennane 2015).

A challenge to improving supports for youth receiving SSI is implementing new interventions in a fragmented service system. Large variation exists across localities in the available education, health, and other rehabilitation supports (NASEM 2018). Moreover, there is no single entry point to obtain those services. Families must navigate a complex and fragmented system to obtain supports that can have conflicting incentives for pursuing activities such as work (Hirano et al. 2018; GAO 2012c, 2017).

The Social Security Administration (SSA) has been working with other agencies to identify strategies to deliver transition services and supports to improve the outcomes of youth receiving SSI. This cross-agency interest has emerged over time as

[1] In 2021, the federal maximum SSI payment is $794 per month, with 23 states providing an optional supplemental amount (The Policy Surveillance Program, n.d.).

the SSI caseload has grown. For example, SSA partnered with multiple government agencies to support the Promoting Readiness of Minors in SSI (PROMISE) demonstration, representing the largest-ever demonstration involving youth receiving SSI. This demonstration, along with other federal policies, has enhanced the focus on serving transition-age youth with disabilities. A key example is the Federal Partners in Transition task force, established to identify strategies to strengthen interagency policy and service coordination for youth with disabilities.[2]

This chapter reviews the findings from SSA demonstrations and other related initiatives to inform options for improving the transition and adult outcomes of youth receiving SSI. As a starting point, we provide an overview of the SSI program rules and characteristics of youth receiving SSI that might influence the youth's and family's choices. We then describe the factors that could affect the youth's and family's human capital development and employment decisions. Next, we summarize findings from evaluations of interventions designed to support youth receiving SSI and other related populations, and discuss how the findings offer lessons for program and policy implementation. We then identify areas for future learning where more evidence is needed to strengthen services and programmatic strategies. In the final section, we offer concluding thoughts about the key lessons learned from our review.

CURRENT PROGRAM RULES

The SSI eligibility criteria have evolved over time since the program's inception in 1974. These criteria provide important context for understanding the outcomes of youth receiving SSI. The eligibility changes are notable because they have contributed to increases in SSI caseloads for children. Moreover, the eligibility criteria include strict medical, income, and asset criteria that can influence the transition decisions of youth. As a starting point, we provide a summary of changes in eligibility and caseload size for children who receive SSI. We then describe the SSI eligibility rules for children and adults and highlight key characteristics and outcomes of youth receiving SSI.

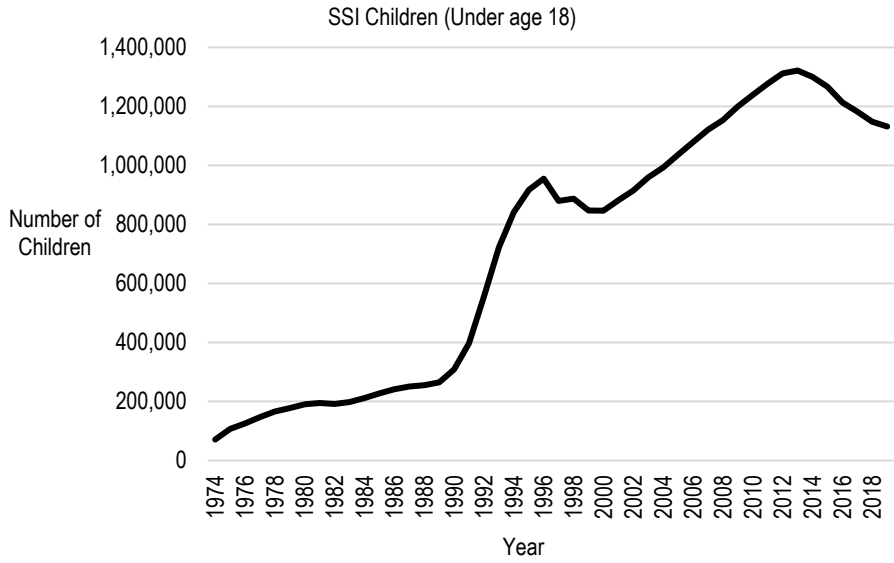
SSI Caseload Size

The eligibility rules for SSI children have changed substantially over time. The SSI program began in 1974 and served a modest child caseload through 1989 (approximately 264,000 children). Following a series of legal and policy changes from 1989 to 1995 that expanded eligibility, the SSI caseload grew by more than 300 percent to about 900,000 children (Berkowitz and DeWitt 2013; Davies, Rupp, and Wittenburg

[2] For more details on the Federal Partners in Transition, see https://www.dol.gov/agencies/odep/program-areas/individuals/youth/federal-partners (accessed May 2, 2021).

2009; Wittenburg 2011).[3] In part due to this growth, the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 substantially revised the child SSI eligibility criteria and required SSA to conduct an eligibility redetermination at age 18. The changes resulted in caseload declines through 2000 (Exhibit 6.1).

Exhibit 6.1. Caseload Trends for Children (Ages 0–17) Receiving SSI



Source: SSA (2019a).

Although the SSI eligibility rules for children have not changed since 1996, there was a steady increase in caseload through 2013. This increase is in contrast to marked declines in other programs' caseloads, such as Temporary Assistance for Needy Families (Schmidt and Sevak 2004, 2017). The factors driving the child SSI program's growth are not well understood, but likely include increases in the number of children living in low-income families, changes in state cash assistance programs, and increasing awareness of childhood disability (Aizer, Gordon, and Kearney 2013; Schmidt and Sevak 2017; GAO 2012c).

---

[3] A series of policy changes led to revisions to the medical eligibility criteria. First, SSA modified the section of the Listing of Impairments that addressed eligibility for children with mental disorders, moving toward a standard based on functional capacity. Second, in its 1990 *Sullivan v. Zebley* decision, the US Supreme Court decided that SSA's listing-only approach for determining disability in children did not reflect the comparable severity provision of the Social Security Act. The Court ordered SSA to assess children individually, which resulted in SSA regulations to implement an individualized functional assessment to determine whether a child could function "independently, appropriately, and effectively in an age-appropriate manner."

Since 2013, the number of children participating in SSI has declined, and applications dropped sharply during the COVID-19 pandemic. In December 2020, some 1.1 million children participated in the program, down from its peak of 1.3 million children in December 2013. There has been a marked decline in child SSI awards during the COVID-19 pandemic (SSA 2021), which could lead to further declines in future SSI caseloads. An important factor likely driving the decline is the closure of SSA field offices during the pandemic.

SSI Eligibility Requirements Differ for Children and Adults

The SSI child eligibility requirements that apply before age 18 differ from the adult requirements starting at age 18. Prior to age 18, youth and their families must meet the medical criteria for children, and the portion of the family's income deemed to the child must be below the SSI income threshold. At age 18, children receiving SSI who wish to continue receiving benefits must undergo an assessment in which SSA determines if they meet the adult SSI eligibility requirements (referred to as the age-18 redetermination). A large share of children ultimately lose their SSI payments and access to Medicaid through SSI because they do not meet the adult eligibility requirements.

Child Eligibility Criteria

The SSI eligibility requirements for children include medical criteria to assess a child's functional capacity. SSA obtains information from medical sources to assess whether a child has

> a medically determinable physical or mental impairment, which results in *marked and severe functional limitations* (emphasis added), and which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than twelve months. (Section 1614. [42 USC § 1382c] (a)(3)(C))

The eligibility requirements also include a resource test whereby SSA deems a portion of the parents' income and assets to the child, including earnings, to determine the child's eligibility.

Once eligible, SSA makes a payment for the child to a representative payee. Typically, the representative payee is a parent or other family member. The representative payee's priority is to meet the child's basic needs (e.g., food and shelter). Additionally, there are provisions that the representative payee seek treatment for the child's medical condition when necessary. SSA has a process for annually reviewing how representative payees use funds, though parents are exempt from this review. In general, once approved, there is minimal oversight of representative payees by SSA (Social Security Advisory Board 2016).

SSA conducts medical continuing disability reviews (CDRs) to assess the child's ongoing medical eligibility for SSI. There are mandatory redeterminations at age 18 and for low-birthweight babies. SSA also conducts periodic CDRs, though its ability to do so depends on its discretionary funding. The frequency of CDRs affects benefit durations; there is a negative relationship between the duration of benefit receipt and the number of CDRs conducted (Hemmeter et al. 2021). This relationship is important because it likely contributes to caseload changes through the fluctuations in discretionary funding for CDRs. Since 2015, Congress has substantially increased SSA's discretionary funding to conduct CDRs (SSA 2019a), and the child SSI program has experienced declining caseloads since that time (Exhibit 6.1).

Children must also continue to demonstrate that they meet the income and asset criteria by reporting any changes to their own or their parents' income and assets. The rules for reporting income are like those of the adult SSI program.

The SSI program has provisions to encourage youth to work, though few report any earnings to SSA. Less than 1 percent of children ages 14–17 receiving SSI reported earnings to SSA in 2017 (Honeycutt, Wittenburg, Crane, et al. 2018). Qualitative evidence suggests that youth receiving SSI, and their families, struggle to understand the program provisions governing earnings, including the special rules that allow youth to exclude earnings from their benefit calculations (Hernandez et al. 2006). For example, under the Student Earned Income Exclusion, SSA excludes earnings up to certain amounts ($1,930 per month, up to $7,770 per year in 2021) in computing the SSI payments for youth under age 22 who regularly attend school. The GAO (2017) found that less than 2 percent of youth benefited from this provision during 2012–2015. The complexity of the SSI rules regarding earnings creates potential challenges for youth and their families in making informed decisions about working and understanding the implications for their SSI and other benefits.

The SSI program also has a provision (Section 301, which applies to all continuing disability reviews) that allows youth to retain SSI payments after age 18 regardless of the age-18 redetermination outcome, though usage is limited. To qualify for Section 301 benefit continuation, a youth must have an approved plan for ongoing participation in services that will enhance employment, such as Vocational Rehabilitation (VR) or continuing special education services.[4] If eligible, the youth's benefits will continue under Section 301 regardless of the age-18 redetermination until the youth completes or ends participation in the services or SSA determines that the services do not contribute to the youth's long-term self-sufficiency.

Finally, SSA provides informational materials to support youth during their transition through the *Red Book* and annual notices. The *Red Book* is a reference to provide all people receiving SSI and Social Security Disability Insurance (SSDI)

[4] For details on Section 301 eligibility, see Program Operations Manual System: "DI 14505.010 Policy for Section 301 Payments to Individuals Participating in a Vocational Rehabilitation or Similar Program" (Effective Dates: 01/06/2017–Present." Accessed May 10, 2021. http://policy.ssa.gov/poms.nsf/lnx/0414505010.

information about employment support provisions. It outlines eligibility rules and provides links to outside employment support programs, such as Job Corps.[5] SSA also sends notices to children ages 14–17 to identify resources to assist in their transition to adulthood.[6] The notices include information about the age-18 redetermination as well as several other SSA work incentives (e.g., SSI continued payments under Section 301). The notices also include information about how youth can benefit from other programs to support their transition to adulthood.

Adult Criteria (Age-18 Redetermination)

At age 18, the medical criteria change to an adult, work-based definition. The adult criteria assess a person is

> unable to engage in any *substantial gainful activity* (emphasis added) by reason of any medically determinable physical or mental impairment which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than twelve months (Section 1614. [42 USC § 1382c] (a)(3)(A))

Additionally, there is a change in how SSA counts the family resources. After age 18, SSA no longer deems a portion of the parents' earnings and assets to the youth (SSA 2021). This change has implications for parental labor supply decisions because parents' earnings are no longer factored into the eligibility calculation. The parent, however, can remain the representative payee, which has potential implications for sharing the benefit check within the family.

A substantial share of youth do not meet the adult eligibility requirements at age 18, though the rate varies by cohort. Between 2000 and 2015, some 52 to 69 percent of children who received SSI remained eligible after the age-18 redetermination (Hemmeter et al. 2021). One factor contributing to the variation in initial eligibility rates is the variation in the volume of CDRs conducted before the age-18 redetermination. If a youth has not had a CDR until age 18, the probability of remaining eligible for SSI after age 18 is lower than for a comparable youth who passed a prior CDR. Thus, the size and characteristics of the population reaching and passing the age-18 redetermination changes over time with the number of child CDRs completed.

The adult eligibility changes have implications for transition planning and the potential need for intervention supports. As children receiving SSI approach age 18, they and their families might need to plan for a new income source to replace the SSI payment. Additionally, they might need to identify other sources of health insurance should they no longer be eligible for Medicaid. Finally, the change in income

---

5 The *Red Book* (SSA 2020e) is available at https://www.ssa.gov/redbook.

6 The information notice (SSA 2020i) is entitled *What You Need to Know About Your Supplemental Security Income (SSI) When You Turn 18.*

eligibility requirements from a focus on family resources to those of the youth affects financial planning for the youth and family. The family must consider whether a parent or guardian will remain the representative payee and how to use the SSI payment, which can be an especially important issue if the youth needs to use the income to live independently.

Characteristics and Outcomes of Youth Receiving SSI

Analyses of SSA administrative data on a cross-section of youth (ages 14–24) receiving SSI in 2017 provide insights into their characteristics and potential support needs (Honeycutt Wittenburg, Crane, et al. 2018). These youth were predominantly male (64 percent), and most had a primary diagnosis related to a mental impairment (80 percent), which includes intellectual disability, autistic disorders, development disorders, and several other types of disorders (e.g., mood disorders). These characteristics underscore the potential need for impairment-related supports.

Family characteristics are also relevant to transition planning. Data from a 2013 cohort indicate that most children receiving SSI live in one-parent families (71 percent) and with other siblings (74 percent) (Bailey and Hemmeter 2015). Additionally, SSI is a source of income for another member (adult or child) in approximately one-fifth of families.

Another relevant issue in considering supports for youth receiving SSI is the geographic variation in SSI participation. There is a clustering of SSI caseloads for children by state and county, with higher participation rates in northeastern and southern states and lower rates in western states (Wittenburg et al. 2015). Substantial variation also exists in programmatic outcomes across state lines. For example, state age-18 redetermination cessation rates range from 20 to 47 percent, and there is evidence of large cross-state differences in adult program and employment outcomes (Hemmeter, Mann, and Wittenburg 2017). These patterns reflect, in part, geographic variation in characteristics of the population, income, and service differences—areas with high rates of low income also have high rates of youth SSI recipients, particularly in southern and northeastern regions. The variation also reflects geographic differences in other social programs and policies that interact with SSI (Meyers, Gornick, and Peck 2002; Schmidt and Sevak 2017).

The racial composition of youth who receive SSI also varies by geographic region. Overall, about half of youth who received SSI in 2000 were non-White (Wittenburg 2011). SSA no longer publishes statistics on race and ethnicity, so information about more recent cohorts is unavailable (see Martin 2016). Evidence from the PROMISE demonstration implemented in 11 states suggests substantial variation in racial and ethnic composition by geographic location. For example, non-Hispanic Black participants represented from 11 percent (the consortium of six western states) to 62 percent (Maryland) of all enrollees. The variation across states in the percentage of Hispanic participants was similar (8 to 65 percent).

Many former child SSI recipients face challenges in transitioning to adulthood. SSI children experience high dropout rates, unmet health care needs, and low employment rates (Deshpande 2020; Hoffman, Hemmeter, and Bailey 2018; Wittenburg 2011). Former child SSI recipients whose eligibility ceased in adulthood experience greater income volatility later in life, and regardless of cessation, former child SSI recipients have low average lifetime earnings (Deshpande 2016a).

There is descriptive evidence that interventions, such as training and VR services, can enhance outcomes for youth receiving SSI. For example, there is a positive correlation between the use of VR services and the adult earnings of former child SSI recipients (Hoffman, Hemmeter, and Bailey 2018). Similarly, there is evidence of correlations between participation in a private vocational training program, Bridges from School to Work, targeted to urban youth with disabilities and the youth's long-term employment and earnings (Hemmeter et al. 2015). Although the positive correlations are promising, the studies lack a comparison group, which is a key feature of the demonstrations serving youth receiving SSI we review later in this chapter. Aside from the information that SSA provides via the *Red Book* (2020e), SSA does not directly refer youth receiving SSI to specific supports, although it does make known their potential availability; youth and families must proactively identify these supports on their own or with the aid of schools or other programs.[7]

THEORY AND IMPLICATIONS FROM ECONOMIC THEORY

We describe a model of potential determinants of adult outcomes of youth receiving SSI based on theoretical and empirical findings from the literature. This model provides a general framework of factors that influence the outcomes addressed by several of the demonstrations discussed later in the chapter. As a starting point, we review human development and labor supply theory to highlight theoretical factors that influence adult outcomes. We then summarize applications and related literature for youth receiving SSI.

Human Development Theory

Skills and attributes developed during childhood are a factor in determining adult outcomes. Research suggests that at least 50 percent of earnings differences across adults are due to personal characteristics established by age 18 (Huggitt, Ventura, and Yaron 2011). Therefore, parenting decisions and the circumstances during childhood

[7] When Congress enacted the Ticket to Work and Work Incentives Improvement Act of 1999, SSA lost its ability to refer beneficiaries to state VR agencies to avoid giving preferential treatment to those agencies over other participation Ticket to Work providers. More recently, Congress and other stakeholders have shown interest in finding ways for SSA to encourage and facilitate use of VR services, especially for transition-age youth who are not eligible for the Ticket to Work program (GAO 2017).

are key factors in determining human capital accumulation and must be considered in designing interventions to improve youth's adult outcomes.

Human development theory encompasses the youth's development, parental investments in their children, and parenting style. Fundamentally, this theory posits that parents seek to maximize their children's long-term welfare. Parents decide how much time and money to invest in their children's development based on their preferences and expectations, the child's preferences and human capital endowment, and the family's resource constraints. Skill accumulation by the child depends on parental investments and the child's investments, the technology of skill formation, and environmental factors (such as the influence of schools, neighborhoods, and peers); all of these factors can be influenced by parental choice and parenting style.

There are two notable findings of this literature that relate to youth who receive SSI.[8] First, the youth's cognitive and noncognitive skills influence school and labor market outcomes by age 30 (Francesconi and Heckman 2016; Cunha and Heckman 2008). School completion and postsecondary education depend more on cognitive skills (problem-solving, intellect, and memory). Importantly, cognitive abilities and intelligence develop in early childhood and remain relatively stable into the adult years (Campbell et al. 2001; Heckman 2011; Heckman and Mosso 2014). Noncognitive (personality, social, and emotional) skills can continue to evolve from early childhood through early adulthood. The implication is that early childhood interventions should focus on cognitive skill development, and interventions in the youth's adolescence should focus more on noncognitive skills, given that these skills are still developing and amenable to change.

Second, socioeconomic factors play important roles in skill development and parental support. There are disparities in cognitive and noncognitive skills across socioeconomic groups at early ages, with children from disadvantaged families having lower skills throughout childhood relative to children from advantaged families (Cunha et al. 2006; Cunha and Heckman 2007). Numerous studies find that these early life disadvantages and environments affect various later life outcomes, including employment (Almond and Currie 2011; Cunha et al. 2006; Heckman and Mosso 2014).

The socioeconomic issues are particularly relevant given that youth receiving SSI live in families with limited resources and potentially face systemic issues in accessing supports. Parents with fewer resources often have less education than their peers and face greater time and resource constraints in supporting their child's learning. Parents' circumstances are important because parental time inputs are critical to the child's early development, and parents' decisions and attributes affect a child's formation of skills and success in later life. Youth receiving SSI might need more substantial

[8] See Heckman and Mosso (2014) and Francesconi and Heckman (2016) for comprehensive reviews.

investments to compensate for the lack of parental knowledge and skill suggested by the literature as being prevalent among low-income families.

Labor Supply Theory

Labor supply theory models the factors that determine the number of hours individuals will choose to work. In this framework, individuals seek to maximize their well-being by consuming goods and leisure (nonwork). Because goods cost money, they must work to earn money to buy them. Consequently, for those who do not work, the economic tradeoff is to consume more leisure; those who work can consume more goods but must give up some leisure.

Youth Incentives

The age-18 redetermination creates tradeoffs for the youth related to employment before turning age 18. On the one hand, there are incentives to encourage youth to work. For example, the SSI Student Earned Income Exclusion, noted previously, allows youth to make $1,930 per month (up to $7,770 per year, in 2021) without the earnings affecting their SSI payments or eligibility. Additionally, youth can qualify for SSDI based on their work history, creating opportunities to receive SSDI benefits and Medicare coverage. Therefore, the additional income from earnings and the potential for Old-Age, Survivors, and Disability Insurance coverage create incentives for youth receiving SSI to work. Early work experience is also a strong predictor of post-school employment among youth with disabilities (Test et al. 2009; Carter, Austin, and Trainor 2012).

However, there are also disincentives for the youth to work because of the different SSI eligibility rules that apply at age 18. At age 18, youth must meet the work-based, adult SSI eligibility criteria to continue receiving benefits. Consequently, if a youth demonstrates the ability to engage in SGA before the age-18 redetermination, a family might mistakenly believe that it could jeopardize the youth's SSI eligibility as an adult. For example, suppose parents believe it unlikely that the youth will be self-sufficient as an adult and are concerned about the youth maintaining SSI and Medicaid eligibility past age 18. They might discourage the youth from working before completing the age-18 redetermination. Their limiting or preventing the youth's work activity could lead to long-term negative impacts on the youth's employment and self-sufficiency.

The labor-leisure tradeoff and potential work disincentives remain for youth who apply for SSI after age 18. Although various program provisions allow individuals receiving SSI to maintain Medicaid eligibility and keep more of their SSI benefits as their earnings rise, some level of earnings will eventually jeopardize SSI eligibility. Modest and consistent earnings might also make the youth eligible for SSDI. If so, the work incentives provisions become even more complicated because of how SSI and

SSDI interact and the different rules of each program governing how earnings affect payments.

National survey data and qualitative interviews with beneficiaries suggest that a lack of knowledge about the SSA work incentives provisions are common (SSA 2018a; O'Day et al. 2016). The lack of understanding of the regulations combined with a fear of benefit loss might prompt some SSI recipients to limit their earnings. For example, although provisions allow SSI and SSDI concurrent beneficiaries to retain SSI eligibility when working above the SGA threshold, one study found that many employed beneficiaries appear to keep their earnings just below that threshold, presumably to avoid jeopardizing their eligibility for benefits (Schimmel, Stapleton, and Song 2011).

Parents and Other Household Members

The SSI program eligibility requirements also have implications for other family members, especially parents, given the deeming rules. The effects of SSI income on parental labor supply and family income will depend on parents' preferences and behavior (Duggan and Kearney 2007). SSI can increase the total family income if parents do not reduce their earnings. The cash benefit can also affect the labor-leisure tradeoff of parents. The nonlabor income reduces leisure costs and allows parents to invest more time in caring for a child receiving SSI. It can also create a disincentive to work with no increase in time invested in the child.

Several studies examined how child SSI payments affect parental labor supply, child well-being, and siblings' earnings. Findings on the effects of child SSI receipt on parental labor supply are mixed. Some studies found no relationship between SSI receipt and parental earnings (Duggan and Kearney 2007; Hemmeter 2015); others found a negative relationship (Deshpande 2016b; Guldi et al. 2018). Studies also show that SSI receipt is associated with improved child outcomes (Guldi et al. 2018; Ko, Howland, and Glied 2020). Finally, there is some evidence that SSI income influences the income of other family members. Deshpande (2020) found that SSI child income supports the SSI child's siblings' future adult earnings. This effect likely occurs through maintaining the overall family income and other resources for the family. As noted previously, the child SSI payments represent a large share of total family income for families receiving the payments.

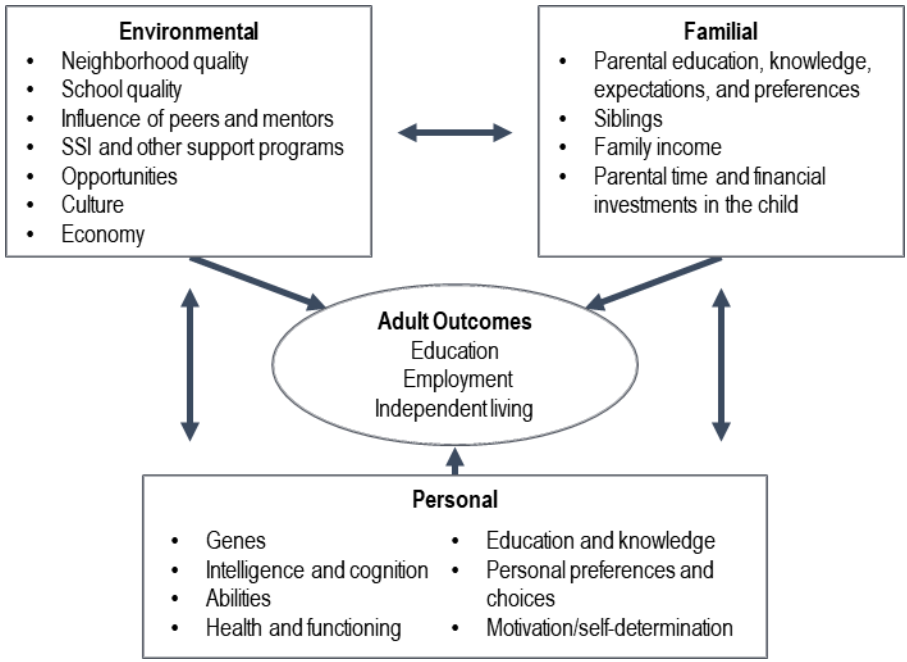
In theory, parents of youth receiving SSI could choose not to invest in their child or could undertake activities intended to ensure that the child continues to meet the SSI medical eligibility criteria to maintain benefits. We are unaware of any evidence supporting this premise. However, the strict resource limits for SSI eligibility might create disincentives for some families to save, which could affect their ability to invest in their children if they lack access to credit or are unaware of mechanisms, such as Achieving a Better Life Experience accounts and SSI's Plan to Achieve Self-Support, that excludes assets used for child investments from SSI eligibility considerations.

Because SSI targets families with limited means, any negative effects of SSI on parental savings and investments in the child are likely limited.

Summary of Factors That Influence Outcomes of Youth Receiving SSI

The above discussion suggests that the youth's personal, family, and environmental factors play an important role in influencing adult outcomes (Exhibit 6.2). The many factors that influence child development and adult success suggest that affecting outcomes of youth receiving SSI is a complex undertaking. There is likely a great deal of heterogeneity across youth and at various ages regarding their needs and circumstances. Focusing on only one or a small set of factors is unlikely to result in meaningful long-term impacts.

Exhibit 6.2. Determinants of Adult Outcomes of Youth Receiving SSI Suggested by Theory and Literature



The various factors influencing youth outcomes suggest numerous ways in which policy and practice might improve outcomes of youth receiving SSI. A large body of literature supports methods effective in helping transition-age youth with disabilities achieve better employment and independent living outcomes as adults. These practices seek to address many of the factors shown in Exhibit 6.2. This literature has been compiled by the National Technical Assistance Center on Transition (NTACT). The NTACT matrix (2016) summarizes the evidence on the practices and predictors that

affect the outcomes of transition-age youth with disabilities. Similarly, the National Collaborative on Workforce and Disability for Youth also has compiled evidence on effective transition practices that form the basis for its framework for successful youth transition, called *Guideposts for Success* (NCWD/Y 2019).

NTACT's matrix and the National Collaborative on Workforce and Disability for Youth's *Guideposts for Success* encompass secondary school practices, VR practices, and predictors of postsecondary outcomes related to education, employment, and independent living. The predictors and practices with research evidence linking them to youth outcomes are too numerous to list here. Still, they encompass a wide range of factors related to school curricula, school transition planning, autonomy, self-determination, life skills, academic achievement, career preparation, and work-based learning experiences, as well as support service delivery, cross-agency collaboration and connectivity, parent expectations, family involvement, and work incentives planning and benefits counseling. SSA does not have purview over many services noted above and so would need to collaborate with another entity to expand services to youth receiving SSI. Nonetheless, several have been incorporated in SSA's demonstrations that targeted youth with disabilities, which we describe in the next section.

DEMONSTRATION FINDINGS AND POLICY PROPOSALS

As a starting point to identify intervention lessons, we review four demonstrations that include youth receiving SSI as a target population: Structured Training and Employment Transitional Services (STETS), Transitional Employment Training Demonstration (TETD), Youth Transition Demonstration (YTD), and PROMISE. These demonstrations are notable for their rigorous designs, implementation scale, intervention approaches, and focus on youth with disabilities. The US Department of Labor (DOL) funded STETS in 1981. SSA used its authorities under Section 1110 and Section 234 of the Social Security Act to support TETD, YTD, and PROMISE.

To identify additional lessons, we draw on interventions implemented for other populations of youth and young adults. We review evidence from interventions that served youth with disabilities and low-income families, given these populations' overlapping characteristics with youth receiving SSI. We also review evidence from studies of interventions that targeted young adults, including interventions described in other parts of this book. Finally, we summarize ongoing interventions and policy proposals. These initiatives represent areas where new evidence could emerge in the future.

Demonstrations Focused on Youth Receiving SSI

The implementation areas, target populations, and intervention services of the four demonstrations focused on youth receiving SSI have evolved over the last 40 years. The earliest two demonstrations (STETS and TETD) began in the 1980s and

included transitional employment supports for youth with intellectual disability. The other two experimental demonstrations, implemented since 2006 (YTD and PROMISE),[9] provided intervention services to a broader group of youth. The broader scope of the activities since 2006 reflects an interest in understanding how to provide supports to the growing and changing composition of SSI caseloads. A common theme across all four demonstrations is the strong emphasis on employment services. Below we provide an overview of each demonstration and summarize its key findings.

STETS: Transitional Employment Supports for Young Adults with Intellectual Disability

STETS Findings

- *Impacts (22 months):* Increased employment, earnings, and income; no impact on disability benefits
- *Costs:* $8,800 per participant ($24,059 in 2020 dollars)
- *Key findings:* Established the efficacy of transitional work services

DOL funded the STETS demonstration in 1981 to test the delivery of transitional work supports to youth with intellectual disability. Social service agencies recruited and randomly assigned 467 youth, most of whom received SSDI or SSI at ages 18–24 in five cities. The STETS services included three phases of work supports (job exposure, on-the-job training, and post-employment follow-up). On average, treatment enrollees received 11 months of services (Kerachsky et al. 1985). The study tracked youth at several intervals up to 22 months following enrollment (Kerachsky and Thornton 1987).

The STETS evaluation established the short-term effectiveness of transitional employment supports in increasing employment and earnings (Kerachsky and Thornton 1987; Kerachsky et al. 1985). The intervention increased employment and earnings at 15 and 22 months after enrollment. Of note is that STETS treatment group members were substantially more likely than their control group counterparts to work in competitive jobs and less likely to work in sheltered workshops. The employment impacts (12 percentage points) were large relative to the control group's 19 percent employment rate. The treatment group earned $16 per week more than the control group. The intervention also led to increased income. However, the employment and earnings effects were not large enough to allow the youth to live independently or replace disability benefits. The study detected no differences in SSI or SSDI receipt or amount at 22 months between the treatment and control groups.

The demonstration findings underscored the effectiveness of transitional employment supports in competitive employment for youth with significant disabilities (Kerachsky and Thornton 1987). This finding is notable given that many control group youth were working in sheltered employment settings. The findings added to the growing descriptive literature at the time demonstrating the viability of

[9] YTD also included some smaller, non-experimental projects that began in 2003 (Martinez et al. 2010), which we do not describe.

competitive employment as an alternative to sheltered work. The STETS demonstration, and subsequently TETD discussed below, provided strong evidence that competitive employment was a realistic goal for youth with intellectual disability.

TETD: A Bigger Version of STETS

SSA funded TETD as a follow-up to the STETS demonstration to test customized transition supports to a larger population of youth and young adults receiving SSI at ages 18–40. Like STETS, TETD focused services on youth with intellectual disability. Unlike STETS, TETD service providers used SSA administrative lists to recruit youth receiving SSI with intellectual disability into the study. TETD represents the first of several youth demonstrations to use SSA administrative records to recruit youth participants.

TETD Findings

- *Impacts (up to 72 months):* Increased employment, earnings, and income; reduced disability benefits
- *Costs:* $5,600 per participant in 1987 ($13,016 in 2020 dollars)
- *Key findings:* Reinforced the importance of transitional supports. Larger impacts for subgroups with higher IQs and more likely to be living independently

Enrollment in TETD began in 1985 and services were provided through 1987. Of the 13,800 eligible SSI recipients invited to participate in TETD, 745 (5.4 percent) enrolled. Relative to STETS, TETD had a larger sample and implementation area (13 demonstration communities) and a more extended follow-up period (up to 72 months) (Decker and Thornton 1995).

TETD employment services were like those in STETS, though TETD put a greater focus on transitional (time-limited) supports. The average length of service receipts varied between 6 to 18 months (Prero and Thornton 1991). The specific TETD services included time-limited job development, on-the-job training, and postplacement services. The demonstration also included waiver exclusions for any income earned from a job obtained through it.

TETD documented several qualitative findings related to service delivery and participant perspectives. Intervention providers faced challenges in convincing employers and family members of the benefits of transitional supports and getting youth needed transportation options (Prero and Thornton 1991). For those participants who received services, however, the intervention shifted costs away from expensive sheltered employment, resulting in savings that could offset the TETD intervention costs. Additionally, the evaluation cited favorable qualitative effects on the youth, such as enhanced quality of life, better social interaction, and higher self-esteem.

The TETD evaluation confirmed that transitional services led to increased employment, but not enough for earnings to completely replace SSI payments (Decker and Thornton 1995). The pattern and size of the impacts were like the STETS demonstration. For example, the cumulative impact on earnings over 72 months was $4,300 (not adjusted for inflation), representing a 72 percent increase. Unlike STETS,

TETD also resulted in a reduction in SSI payments. The decline was modest: just $870 over the study period. Treatment group participants increased their total family income, but their earnings did not completely replace their SSI payments.

The TETD impacts varied by site and subgroup, underscoring the importance of customizing services to meet youth's specific needs. For example, the employment impacts were larger for youth with higher IQ scores and those living independently. Moreover, the TETD programs that provided customized supports had larger impacts than programs that did not attempt to customize services.

A limitation of TETD (and STETS) was that it was difficult to generalize the findings to a broader set of programs and policies. This challenge reflects the intervention's rollout with a sample of volunteers who participated at relatively low rates. As a result, it was unclear whether the intervention would result in similar effects in other areas. Nonetheless, the evaluation findings provide insights into the potential for interventions to improve the outcomes of youth with intellectual disability, which at the time represented a large portion of youth receiving SSI (some 30 to 40 percent).

YTD: Services Delivered to a Broad Population of Youth with Disabilities through Service Providers and SSI Program Waivers

The foundation for funding a larger project involving SSI youth started from the Youth Continuing Disability Review project conducted by Maximus. This project included SSI youth in Maryland and Florida ages 15 and 16 who had a CDR. The project provided youth with access to services on skill assessments, career aspirations, educational goals, health care needs, reasonable accommodations, employment supports, and community and governmental transition supports. The study findings emphasized the importance of individualized strategies to help youth succeed in the workplace (Maximus 2002). The study noted a major issue in providing services was overcoming difficulties associated with the lack of coordinated services across key stakeholders in the school system who were unaware of many special SSA program rules.

YTD Findings

- *Impacts (up to 120 months):* In some programs, impacts on any earnings diminished over time. Income and SSI benefits increased, consistent with use of waivers. In some programs, improvements in some social outcomes, such as reductions in arrests
- *Costs:* Ranged from $5,232 per participant in Erie County to $8,628 per participant in the Bronx
- *Key findings:* The sites with more intensive employment supports tended to have larger employment impacts. In delivering early intervention services, it is important to have well-defined target populations and services. The YTD service costs were generally less than costs of TETD and STETS, though those two had more limited impacts. The duration of services is also important and highlighted considerations for how to sustain programs beyond the demonstration to enhance both outcomes and options to serve youth.

SSA funded YTD programs to test the delivery of employment and other services with waivers to broad target populations of youth with disabilities. The first enrollment into YTD programs began in 2006. The full evaluation tracked outcomes for each project annually for three years (Fraker, Mamun, et al. 2014; Fraker et al. 2018). Additionally, a follow-up study examined outcomes for up to 10 years after enrollment using administrative data only (Hemmeter and Cobb 2018).[10]

The YTD service components followed a modified version of the effective practices outlined in *Guideposts for Success*.[11] The customized services in *Guideposts for Success* addressed the need for individualized services highlighted in the Youth Continuing Disability Review project. The features included work-based experiences, youth empowerment activities, family involvement, system linkages, and benefits counseling. The YTD services emphasized the work-based supports, given their importance in improving employment outcomes as identified in the literature. The intervention also included waivers that modified SSI program rules related to reporting income, the age-18 redetermination, and CDRs. For example, one of the waivers removed the age restriction on the student earned income exclusion.

YTD included several organizations that led service implementation (including private providers, non-profits, and a university) to target populations in six sites: Colorado, Florida (Miami), Maryland, New York (Erie County and the Bronx), and West Virginia.[12] Five YTD programs served youth receiving SSI; one program (Maryland) focused on serving youth at risk of SSI entry. In total, the six YTD programs enrolled 5,103 youth (Fraker, Mamun, et al. 2014). The implementation and evaluation scope allowed for analyses of a broader population of youth receiving SSI residing in a mix of rural and urban areas, compared with the previous youth demonstrations.

---

[10] Hemmeter and Cobb (2018) estimated 8-year earnings impacts for all YTD programs and estimated 10-year impacts for a subset of programs.

[11] The framework was based on a modified version of the *Guideposts for Success*. The YTD program and technical assistance teams adapted *Guideposts for Success* to meet the needs of youth receiving SSI (e.g., by adding benefits counseling), though retained the emphasis on the importance of work-based experiences identified in prior studies (see Luecking and Wittenburg [2009] for more details).

[12] SSA ultimately selected six programs for implementation of larger-scale interventions (Fraker and Rangarajan 2009). SSA selected the programs based on proposals and a pilot, where the evaluation team reviewed and made recommendations for how each program could implement and scale its interventions. The six programs had latitude to serve youth receiving or at risk of receiving SSI between the ages of 14 and 25 using service models that fit the *Guideposts for Success* model. Additionally, the evaluation provided technical assistance to programs to support the implementation according to the *Guideposts for Success* model. YTD also included non-random assignment programs in other states (see Martinez et al. [2010] for more details). Camacho and Hemmeter (2013) summarizes findings on service receipt and outcomes from two of the non-random-assignment programs, including detailing the experience of one youth.

The YTD evaluation team used SSA administrative data and worked locally in partnership with the programs to recruit and enroll participants. The onsite work with partners was essential in the intervention to families and building trust in the effort. YTD program and evaluation staff reported that waivers were a strong inducement for youth to enroll, underscoring a critical service component. The enrollment rates ranged from 16 to 30 percent across programs (Fraker, Mamun, et al. 2014).

Nearly all YTD youth received some services, though the intensity of services, particularly employment services, varied by program (Fraker et al. 2018). As one example, service delivery ranged from 7 to 43 hours. The three programs with the most considerable employment impacts also had the most employment service hours. A technical assistance team monitored service delivery and used metrics on the type and amount to support program staff in delivering services. This technical assistance helped program staff provide consistent services with a focus on employment during the demonstration. The evaluation noted that sharpening the focus on employment service supports could be beneficial to other service providers.

In all programs, the YTD interventions increased the likelihood of employment service use. Despite the increase, YTD did not increase the total hours of service use across all providers (YTD or non-YTD). Thus, there appeared to be some substitution of participation in YTD services (focusing on employment) away from non-YTD services.

The estimated employment impacts varied by program and diminished over time (Hemmeter 2014; Fraker, Mamun, et al. 2014; Hemmeter and Cobb 2018).[13] In year one, three programs increased employment (the Bronx, Florida, and West Virginia). The impacts in two of the programs (the Bronx and West Virginia) were initially large (16 and 24 percentage points, respectively) compared with impacts in later years. In part, the large impacts represent aspects of YTD services that included employment as an extended part of services, especially in the Bronx program, which offered summer youth employment programs. The third program (Florida) had relatively modest employment impacts in year one (6 percentage points). In the second year, two programs (the Bronx and West Virginia) continued to sustain employment gains. The impacts decreased from the year one estimates (6 and 8 percentage points, respectively) (Hemmeter 2014). In year three, two programs (Florida and West Virginia) continued to have employment gains on the order of 6 to 8 percentage points

[13] The evaluation included measures of *any employment* from the survey, *any earnings* from administrative records, and *earnings levels* from earnings records. For simplicity, we summarize the *any earnings* outcomes here given they are available for all years whereas the survey findings are available only in years one and three. Some programs had impacts on *any employment* from the survey, but not administrative records (e.g., Erie County in year three; see Fraker, Mamun, et al. [2014]). However, the point estimates are all below 8 percentage points, so there is no substantive difference in the broad interpretation between the *employment* and *any earnings* measures shown here.

(Fraker et al. 2018). However, no program had an impact on employment after year three (Hemmeter and Cobb 2018).[14]

The evaluation cited a strong relationship between service intensity and employment impacts for two programs that generated employment impacts through year three (Florida and West Virginia). These two programs were also the only programs that had impacts on "productive activities," which included participation in employment, education, and training. The evaluation noted that these two programs also had intensive employment service delivery interventions that differentiated them from the other programs (Fraker et al. 2015; Fraker et al. 2018). Conversely, the programs that did not have as strong a focus on employment were less likely to generate impacts on employment and other productive activities.

The five programs that included youth receiving SSI support produced sustained increases in SSI benefit amounts, which increased income. The cumulative impacts on SSI benefits ranged from about $3,000 to $6,000 in the seventh year after enrollment (Hemmeter and Cobb 2018). These benefit increases are not surprising given that the waivers offered under YTD provided protections for income and from eligibility redeterminations that increased benefit duration.

The findings of the sixth program (Maryland), which did not generate employment or benefit impacts for at-risk youth, provide insights into the challenges of providing early intervention services. The program offered intensive services, but qualitative findings documented that the counterfactual service environment was already strong. The youth in this program used YTD services to supplement existing supports that were already available. The lack of impacts likely reflects that other similar supports were available in the area. This finding further underscores the importance of developing customized approaches that fill a specific need among well-targeted populations.

Another notable finding was that two YTD programs (the Bronx, Florida) achieved reductions in youth arrests, though one program (Colorado) increased arrests.[15] The results are notable given the relatively high arrest rates among young adults with disabilities relative to those without disabilities (Wittenburg 2011). The evaluation could not specifically identify the components of the interventions that generated these results, though it noted large service differentials between these three programs that could influence impacts (Fraker, Mamun, et al. 2014). The programs with more intensive services generated larger impacts. The evaluators hypothesized

---

[14] The Hemmeter and Cobb (2018) results are based on unpublished slides. The findings reported that employment increased in year 6 for West Virginia and in year 10 for the Bronx, with effect sizes of 6 to 7 percentage points. All other results were not statistically significant. The authors concluded there were limited sustained impacts on employment after the intervention period, like effects in other training programs.

[15] Regarding other primary outcomes, no program had impacts on youth self-determination. The final evaluation report noted that the methods for measuring self-determination were limited at the time of the YTD evaluation (Fraker, Mamun, et al. 2014).

that, as with intensive employment supports, well-designed and intensive other services might support youth in reducing contact with the justice system.

In summary, the findings show the potential for intensive services and waivers to improve the outcomes of broader populations of youth receiving or at risk of receiving SSI than those served by previous demonstrations. Well-designed and targeted interventions generally led to promising impacts, particularly in the demonstration's first few years. The YTD service costs were substantially lower than the costs of its STETS and TETD predecessors, though YTD's impacts were also more limited and diminished more substantially over time. These findings raise the important issue of how to determine the optimal intensity and duration of services needed for youth receiving SSI to succeed.

PROMISE: Supports Delivered through State Agencies to Children Receiving SSI and Their Families

Beginning in 2013, PROMISE tested state-based intervention services delivered to a large sample of children receiving SSI who were age 14 to 16 when they enrolled in the study. The ongoing PROMISE evaluation measures impacts for a wide range of youth and family outcomes at 18 months and five years after enrollment. To date, 18-month impact findings are available; future evaluations will include five-year impact estimates (Fraker, Carter, et al. 2014; Mamun et al. 2019).

PROMISE Findings

- *Impacts (18 months, long-term impacts forthcoming):* Eighteen months after enrollment, each program increased youths' use of transition services and family members' use of support services. None increased youth school enrollment, but all had increased youth's receipt of job-related training and employment. Four programs increased youth's earnings and total income, but only one reduced youth's federal disability payments. One program increased parents' receipt of education and training, but none affected parents' employment, earnings, or income.
- *Costs:* Annual cost per enrollee ranged from $5,490 to $9,148 across programs.
- *Key findings:* Providing services to families in addition to youth has the potential to improve youths' outcomes; however, there are challenges to engaging families and sustaining family-focused services.

Five features differentiate PROMISE from the youth demonstrations described above (Fraker, Carter, et al. 2014). First, the US Department of Education (ED) funded and provided oversight for the implementation, and SSA funded the evaluation. ED funded six PROMISE programs that encompassed 11 states.[16] Second, state agencies led the implementation of services and were required to engage in cross-agency collaboration on services

[16] Two PROMISE projects were implemented statewide (Maryland and Wisconsin), three were implemented in selected geographic areas of a state (Arkansas, California, and New York), and one included a consortium of six western states that implemented PROMISE statewide in each (Arizona, Colorado, Montana, North Dakota, South Dakota, and Utah).

provision. Third, PROMISE services targeted a generally younger population of children receiving SSI (ages 14–16) than the populations targeted by the previous demonstrations. Fourth, PROMISE had a much stronger focus on providing services to family members as well as the youth, particularly on providing parent training and education.[17] Finally, PROMISE included more youth receiving SSI (N=13,444) than any other SSI demonstration.

PROMISE recruitment began in 2014. SSA provided the PROMISE programs with lists for recruitment of eligible youth receiving SSI residing in the programs' service areas. The programs had to enroll at least 2,000 youth and their families over two years. The evaluator worked with the programs to randomly assign youth to either the treatment group or the control group at enrollment. Using various methods, including phone, mail, and in-person outreach and incentive payments, the six programs enrolled a total of 13,444 youth, representing between 16 and 43 percent of the eligible youth they contacted (Livermore et al. 2020). The enrolled youth were generally representative of the broader population of youth receiving SSI in the catchment areas.

The PROMISE programs delivered services for approximately five years. ED required the programs to deliver four core services at a minimum: (1) case management to youth and their family members, (2) benefits counseling and financial education, (3) career and work-based learning experiences for youth, and (4) parent training and information to help parents support and advocate for their youth, as well as resources for improving the education and employment outcomes of the parents themselves. Case management was the cornerstone of the intervention, used to identify youth and family needs and connect them to services and information that would improve their education, employment, and self-sufficiency. Each program developed collaborations with existing public service providers, including the state VR, Medicaid, and developmental disability agencies; high schools; workforce centers; Work Incentives Planning and Assistance projects; and independent living centers. Service providers encouraged youth to apply for services and supports, such as Section 301, that might benefit them. These collaborators served on project advisory committees and, to varying degrees, partnered with the PROMISE programs to provide services. The service arrangements varied from formal contracts with public and private providers to less formal referral arrangements with existing services.

All programs experienced challenges in getting some of their services in place (Anderson et al. 2018; Honeycutt et al. 2018; Kauff et al. 2018; Matulewicz, Katz, et al. 2018; McCutcheon et al. 2018; Selekman et al. 2018). The challenges arose from the need to identify providers of some services or develop them from scratch. In most areas, the services did not exist in the community or were not of a scale across the catchment areas to serve all PROMISE enrollees.

---

[17] YTD included family involvement components of services, though the emphasis on this type of service was substantially less than in PROMISE.

The programs also faced challenges in convincing parents to participate in services. For example, some parents viewed the program as being for their youth and were less willing to participate in services for themselves. Relatedly, some parents believed their children were too young to engage in services related to employment. Finally, some were preoccupied with addressing crises arising from their limited income or their child's or their own poor health. These factors limited the time and energy of some parents to engage in PROMISE services.

As of 18 months after enrollment, the PROMISE programs had increased youth use of transition services (Mamun et al. 2019). Most control group youth received some transition services. Nonetheless, the PROMISE programs generated impacts ranging from 27 to 69 percent greater than the control group means for specific types of services. The largest impacts were on PROMISE's core services, which were generally the least-used types of services among control group youth and families. The most common services used by control group youth were transition planning services and life skills training. These represent services that most special education students are likely to receive in the ordinary course of attending high school. However, services that might be more applicable to youth receiving SSI and their families (and the focus of the PROMISE demonstration), including case management, work-based learning experiences, benefits counseling, and financial education, were less commonly accessed under the status quo.

All programs increased the likelihood that youth received job-related training and engaged in paid employment. The employment impacts were substantial for some programs, ranging from 26 to 184 percent greater than the control group means. The programs with larger impacts had contracts with providers or hired dedicated staff to deliver employment-related services and offered wage subsidies. The programs with smaller effects relied more on referrals to existing employment services. These findings underscore the importance of proactively engaging and funding providers to enhance impacts. Notably, the program that generated the largest impact on employment and earnings had the highest cost per enrollee.

Four of the programs (Arkansas PROMISE, CaPROMISE, MD PROMISE, and WI PROMISE) increased youth's earnings and total income, but only one (CaPROMISE) reduced federal disability payments. One program increased the likelihood that the youth had health insurance (from any source) by one percentage point.

None of the programs affected youth expectations, self-determination,[18] or the number of months enrolled in Medicaid. Although the evaluation found some statistically significant differences across subgroups of youth defined by sex,

[18] Self-determination is a concept that encompasses attitudes and abilities that lead individuals to set goals and take actions toward achieving them. The PROMISE 18-month evaluation assessed autonomy, psychological empowerment, and self-realization—three of the four subdomains of the ARC Self-Determination Scale (Wehmeyer 1995) using youth's responses to the 18-month survey.

impairment, and age, there were no consistent subgroup differences across the programs.

All PROMISE programs increased other family members' use of support services. Examples of services included the same core services offered to youth (case management, benefits counseling and financial education, and employment-promoting services) as well as parent training on the youth's disability.

No program had a statistically significant effect on parental employment, earnings, or income, and only one generated an impact on training. The lack of impacts on parental employment and earnings might reflect the challenges noted above in serving parents. Nonetheless, the evaluation found a favorable relationship between youth outcomes and family service use (Levere et al. 2020). This finding suggests that family engagement might favorably affect youth outcomes by increasing youth participation in services as well as through other indirect means.

The early findings suggest that PROMISE services fill essential gaps in services. For example, some services, especially benefits counseling and financial education, were not widely available in the community. The promising early impacts also indicate that the services meet short-term needs. Despite the potential need for services, it remains challenging to engage families in ways that make even larger potential impacts possible. The low-income families whose children receive SSI experience regular crises related to the youth's health condition or the family's limited resources. These crises can disrupt families' and case managers' focus on the ultimate goals of interventions such as PROMISE. Moreover, some parents believed that PROMISE was for their youth and not themselves, which likely limited the programs' ability to address fundamental family issues that could affect the long-term outcomes of the youth. The link between a parent's knowledge, behavior, and circumstances and the youth's outcomes might not have been evident to parents except with respect to services that were directly related to the youth (e.g., assistance with guardianship issues).

Another early lesson is that collaboration across agencies is potentially beneficial in addressing fragmentation in existing supports. Formal contracts between entities, including service benchmarks and funding, appeared to be more effective in ensuring that youth received intended services than were more informal collaboration types (Livermore et al. 2020).

Finally, intensive, family-based interventions are challenging to sustain without funding and incentives to support them. Family case management was the central feature of PROMISE and represented the largest service cost; PROMISE case managers' small caseloads (about 30 or fewer families) contributed to the costs.[19] Although often affiliated with a state agency, these case managers functioned independently from any state program. None of the PROMISE states has continued to provide family-focused case management to youth receiving SSI and families offered

[19] In contrast, state VR counselors typically have caseloads of 100 or more.

under PROMISE. Existing programs face challenges in adopting more intensive targeting and case management services due to cost. Additionally, it is difficult to integrate comprehensive family case management services in the existing system because of programs' other priorities and legal mandates to serve individuals rather than families. Staff affiliated with the Wisconsin PROMISE program proposed a means for incorporating this case management in state VR programs (Anderson, Schlegelmilch, and Hartman 2019). Based on the PROMISE experience, others have also proposed interventions for incorporating system-wide, family-focused case management into the transition landscape (Anderson, Hartman, and Ralston 2021; Karhan and Golden 2021).

Evidence from Other Interventions

Several other interventions have offered supports to people who share some characteristics with youth receiving SSI. These populations include other youth with disabilities, youth with limited resources, and adults with disabilities. In this section, we also briefly review evidence from interventions implemented in other countries. Several of these promising interventions could be viable options for improving the outcomes of youth receiving SSI.

Evidence from Other Interventions

- *Impacts:* Targeted interventions with intensive supports show promise for sustaining impacts. Research shows promising impacts for residential models, employer-based supports, sectoral training, population-specific approaches, and models that tie in specific profiling options
- *Costs:* Vary by intervention approach
- *Key findings:* More-intensive supports are associated with stronger outcomes. There is promise in focusing on residential and job sector training initiatives, such as Job Corps and Year Up, which have provided supports unlike those tested in SSA demonstrations. Other promising supports for youth with disabilities have not been rigorously tested and so offer opportunities for future learning

Employer and Residential Interventions for Youth with Disabilities

Descriptive evidence from recent studies underscores the promise of long-term, comprehensive transition supports improving outcomes for youth with disabilities (Honeycutt, Wittenburg, Crane, et al. 2018). Examples include the Maryland Seamless Transition Collaborative Program, Utah Pathways to Careers, and Marriott Foundation Bridges from School to Work. Like the YTD and PROMISE interventions, the programs in the field provide participants with employment services coupled with other services. However, we cannot say whether the transition programs alone influenced the outcomes because these studies lacked valid comparison groups. Hence, these program interventions represent a potential opportunity for developing further evidence, given their promising descriptive evidence.

There is some evidence of the favorable effects of residential and employer-based training interventions on the employment of youth with disabilities. A study of Job

Corps, the largest free residential education and job training program for youth ages 16 to 24, found employment impacts for youth with medical conditions that were larger than the impacts on other youth (Hock et al. 2017). The study also found a decline in SSI participation resulting from Job Corps participation. Descriptive and experimental studies of Project SEARCH, a business-led one-year employment preparation program for youth with disabilities that takes place entirely at the workplace, found that the program increased the rate of employment (Wehman et al. 2014). A common feature of both interventions that differentiates them from the SSA demonstrations described above is the intensive immersion of youth in services delivered in environments (residential facilities and employer sites) that are different from their usual home and school environments. The compelling evidence from Job Corps and Project SEARCH suggests that they might be worthwhile programs to test referrals involving youth receiving SSI.

Intensive and Sectoral Training Interventions for Other Youth Populations

The evidence from interventions tested with other youth populations, particularly job training interventions, indicates a pattern of impacts that are like the SSA demonstrations; program impacts emerge early on but diminish over time (Treskon 2016). Critical components of successful programs include paid work interventions, financial incentives, and service coordination among education, training, and employment supports.

Interventions that provide intensive training and other supports have proven to be effective among low-income populations (McConnell, Perez-Johnson, and Berk 2014). The rationale for offering intensive training along with other supports to youth with disabilities is that these youth might face several challenges in addition to a lack of occupational skills, including low cognitive and noncognitive skills. There are two reasons why these more intensive programs are likely to be effective for youth receiving SSI. First, they allow youth to move out of their usual environments and receive mentoring and socialization that can contribute to job success. For example, in Job Corps, participants talked about their peers' negative influences in their home neighborhoods; Job Corps removed this influence by relocating youth to the residential Job Corps sites. Second, the programs involve full-time immersion in intensive services (youth.gov, n.d.). The full-time immersion reduces the opportunity for other influences, including family financial and other crises, to interfere with the youth's education and job training. Family crises were a common reason why some PROMISE youth were unable to participate in services as intensively as intended (Hall et al. 2020).

The second strand of evidence is emerging about the potential for workforce development that focuses on specific sectors (sectoral job training), such as information technology, influencing long-term outcomes (Arnold Ventures 2020; Fein, Dastrup, and Burnett 2021). The Year Up program is a notable example because it offers findings from a five-year experimental follow-up evaluation. The Year Up

intervention is an intensive year-long program that trains low-income youth for jobs in high-demand fields, such as information technology and financial services. Year Up increased average earnings by 30 to 40 percent ($7,000–8,000) (Fein, Dastrup, and Burnett 2021). There are also other promising findings of sectoral job training programs, though with shorter follow-up periods. Examples of these other programs include Per Scholas, Project QUEST, and Nevada's Reemployment Eligibility and Assessment program (Arnold Ventures 2020).

Population-Specific Approaches for Adults

The lessons from other SSA demonstrations for adults outlined in this book further emphasize the importance of customizing supports to well-targeted populations. The demonstrations without customized supports, such as the Benefit Offset National Demonstration, Accelerated Benefits, and the SSI Work Incentives Demonstration Project, did not produce long-term employment effects (see Chapters 3, 4, and 5 in this volume). Although the overall employment impacts were limited in these demonstrations, younger participants generally had higher employment and participation rates, underscoring the demand for services by this group (see Chapter 7). Another feature of the interventions that targeted adults is that they differ in their service focus. For example, they lacked a strong focus on family, independent living, self-determination, and education compared with youth interventions, reflecting the different presumed needs of adults and youth.

One area where adult-based demonstrations provide some insights into potential youth intervention options is impairment-specific supports. The Mental Health Treatment Study showed the potential to increase employment for people with mental health conditions through a supported employment intervention and by providing other supports tailored to their needs (Frey et al. 2011). Through its Supported Employment Demonstration, SSA is testing the effects of providing supported employment to people with mental health conditions who are denied disability benefits (Taylor et al. 2020). However, the supported employment model has not yet been tested widely and rigorously among youth (Noel et al. 2018), though one YTD site served youth with severe emotional disturbances at risk of receiving SSI or SSDI. The Mental Health Treatment Study findings, along with those of STETS and TETD described above, suggest the potential for approaches that are tailored to other specific impairments among youth populations, especially those that represent increasing shares of the child SSI caseload, such as youth with autism.

International Evidence

Findings from programs for youth in Europe and Latin America underscore many of the experiences and themes discussed above for programs in the United States (Ibarraran et al. 2014; Kluve et al. 2016). Programs that provided multiple types of services that youth and families could choose from were more likely to have an impact

than those that had a more limited set of services. The ability to provide multiple types of services were common features of programs that had the largest effects (Kluve et al. 2016). There was also evidence that profiling individuals to determine their service needs was beneficial. Under profiling, a provider would group participants based on the level of service need as opposed to, say, specific types of services. For example, a provider would group participants based on the number of services they need (low, moderate, and high). This grouping was helpful to right-size services, determine provider incentives for outcomes, and monitor service delivery. The combination of profiling and incentives is important in identifying services to keep youth engaged in the program, as well as in creating alerts to providers when youth drop below expected service engagement.

Ongoing SSA and Cross-Agency Initiatives

As of this writing, SSA and other federal agencies are supporting several initiatives to improve the outcomes of youth with disabilities. Below, we review both ongoing and proposed initiatives that highlight efforts to support youth receiving SSI, particularly during their adult transition.

Current Initiatives and Proposals

SSA

- Enhance service linkages to other agencies (Ohio Direct Referral Demonstration)
- Predictable CDRs, work incentives and other supports (2021 policy proposals)

Other Agencies

- Improved information related to age 18 redetermination (National Institutes of Health)
- Generate ideas for new interventions to serve youth receiving SSI (Department of Labor)

SSA Initiatives to Support Youth

SSA is supporting an initiative to coordinate VR services for youth in the Ohio Direct Referral Demonstration (SSA 2019b). The effort involves a collaboration between the Ohio Department of Disability Determination and the state's VR agency. The demonstration is testing the effect of providing direct referrals to VR for youth ages 18 and 19 who either are applying for SSI or SSDI or are undergoing the SSI age-18 redetermination. This ongoing experimental study involving 750 youth will assess the impact of the direct referrals on youth's use of VR services, employment, and reliance on disability program benefits.

SSA has also proposed policy initiatives in its budget to test other ways to support youth in the future (SSA 2020g). These include supporting youth receiving SSI through work incentives and referrals to other services. For work incentives, the budget outlines an initiative to disregard all the youth's earned income and eliminate income reporting requirements through age 20. Finally, building on the Ohio initiative and other demonstrations, SSA proposes for Congress to restore its authorization to refer transition-age youth receiving SSI (along with all other SSI recipients and SSDI beneficiaries) to VR services to improve those youths' access to employment-related services.

Cross-Agency Initiatives to Support Youth Receiving SSI

Two cross-agency projects include initiatives to support the transition to adulthood for youth receiving SSI. The first is a National Institutes of Health–supported project to provide more information to families nearing the age-18 redetermination about SSI payments' future availability (Deshpande and Dizon-Ross 2020). The planned random assignment study will provide information to families about youths' potential to retain benefits after age 18, along with other informational resources. The information is intended to address inaccurate beliefs about the future availability of SSI for the youth that might lead to underinvestment in the child's health and human capital. The second is a DOL funded project to identify promising interventions to help youth receiving SSI transition successfully to adulthood (Honeycutt, Contreary, and Livermore 2021). The proposed interventions developed by researchers and practitioners include case management and vocational service interventions, youth and family empowerment curricula, a cross-agency data-sharing tool, scholarships for youth receiving SSI, and delaying use of the SGA criterion in the SSI/SSDI disability determination until age 22.

Policy Lessons

We identify policy lessons from the interventions above and discuss how government agencies have used these lessons to refine services and supports. We first discuss demonstration lessons that inform intervention service design, implementation, and evaluation. Next, we describe our understanding of how government agencies have used the findings from prior demonstrations to reform programs and services, in essence, translating the knowledge gained from the research to policy.

Demonstration Lessons

We identify five lessons for future demonstrations. The lessons address topics related to both the design of demonstrations and the implementation of services.

Demonstration Lessons

1. *Design:* Youth service and support needs differ from those of adults
2. *Recruitment:* Local staff and program incentives enhance enrollment
3. *Intervention services and outcomes:* Intensive service models with clear focus on specific outcomes generate larger impacts
4. *Outcomes:* Long-term impacts through expansion in training and employment opportunities that address systemic gaps
5. *Interagency collaboration:* Formal agreements with financial incentives enhance collaboration

1. Design: Youth service and support needs differ from those of adults

A general lesson from our review is that youth have different service and support needs than adults. The differences have important implications

for service design, particularly meeting the multiple needs of the youth and family. Relative to adults, youth have a more extensive set of needs related to their education and cognitive and noncognitive skill development, which continue to evolve as they age. Thus, a larger set of potential outcomes can be precursors to employment, including self-determination, education, and social engagement. Additionally, families, especially parents and guardians, play a crucial role in supporting and making decisions for the youth; their knowledge, needs, and behavior are also highly relevant in considering how to improve the adult outcomes of youth receiving SSI.

2. Recruitment: Local staff and program incentives enhance enrollment

A second lesson is that recruitment can be enhanced when outreach includes a local presence and program incentives. Incentives to participate in programs might be necessary because youth receiving SSI and their families might fear losing benefits as a consequence of participation. Because the SSI payment often represents a central source of income for the entire family, its potential loss can be a major concern. This concern may be exacerbated by families' confusion about program rules and the implications of the rules for youth outcomes. These factors likely contribute to the low participation rates in demonstrations, which are usually well below 50 percent.

The YTD and PROMISE demonstrations achieved high participation rates (ranging from 16 to 43 percent across sites) relative to the other SSA employment-focused youth demonstrations (about 5 percent). In both demonstrations, a local provider presence was cited as helpful in bolstering recruitment compared with mailings and recruitment originating from a central, non-local source. Additionally, the YTD evaluation cited waivers as being key in reassuring participants about the financial benefits of participation. PROMISE did not include waivers; however, the CDR protection provided through a waiver in YTD was provided as policy in PROMISE.

3. Intervention Services and Outcomes: Intensive service models with a clear focus on specific outcomes generate larger impacts

For those who participate in SSA demonstrations, there is recurring evidence that customized, intensive services improve outcomes. The size of the impacts varies by intervention and population, though all previous SSA demonstrations generated employment gains. A feature that led to generally larger effects was the focus on customized and intensive supports. In general, the more expensive interventions in PROMISE and TETD generated larger impacts.

There is also evidence that having a sharp focus on employment in service delivery improved outcomes. For example, in YTD, there was correlational evidence showing that interventions with more intensive employment services had larger employment impacts. In PROMISE, there was evidence of larger employment impacts for programs that contracted for employment services or had dedicated staff to provide

those services compared with programs that relied on referrals to existing employment services in the community.

4. Outcomes: Long-term impacts through expansion in training and employment opportunities that address systemic gaps

The SSA demonstration findings provide mixed evidence on the potential to influence long-term outcomes. The early STETS and TETD findings suggest the potential to affect long-term outcomes beyond the demonstration period. Within YTD, there was also some evidence that longer periods of service delivery (Miami) resulted in more sustained impacts than the more intense, shorter intervention models (e.g., West Virginia). However, even with the longer-duration YTD interventions, none of the projects sustained impacts beyond the demonstration period.

Outside of the SSA demonstrations, there is evidence that residential or sectoral training supports can improve employment and other long-term outcomes. The Job Corps and Year Up programs resulted in larger employment impacts that persisted beyond the intervention period compared with impacts found in the SSA demonstrations. Additionally, there was evidence that Job Corps reduced reliance on SSI.

A theme that connects the long-term impacts in SSA and other demonstrations is they offered opportunities not available in a youth's environment. In STETS and TETD, youth moved from sheltered to competitive employment environments. Similarly, in the Job Corps and Year Up interventions, the youth had opportunities for training and employment supports that might not have been available in their immediate areas. In providing those opportunities, the interventions fill gaps that might limit a youth's advancement.

5. Interagency Collaboration: Formal agreements with financial incentives enhance collaboration

Finally, the SSA demonstrations provide some insights into creating stronger partnerships among service delivery organizations through the use of contracts and tracking systems to support implementation. In YTD, the technical assistance provider tracked service intensity and types to provide feedback to the implementation partners. This quantitative feedback was useful in conversations with partners and front-line staff about balancing service delivery to ensure all participants received the services as intended.

In PROMISE, there were multiple examples of formal or semi-formal contracting arrangements in supporting collaborations that provide possible examples for future interagency collaborations. Providers used multiple types of arrangements (e.g., fee-for-service and lump-sum payments) to specify services, and, in some cases, payments were small. Even token financing demonstrated the PROMISE program's desire to enhance service capacity and provided a basis for developing a formal scope of work

and reporting between two agencies. Lead agencies could use this information to monitor service delivery and, as necessary, make modifications to services, and identify areas for continued communication and collaboration across agencies (Livermore et al. 2020; Nye-Lengerman et al. 2019).

Translating Research to Policy

Translating Research to Policy

1. *SSA programs:* WIPAs increased benefits counseling focus on youth and families
2. *Other agencies:* SSA demonstrations informed WIOA implementation and future policy proposals

We identified two lessons for program and service reforms at SSA and other agencies in service transition-age youth. We identified these lessons based on interviews with key stakeholders at government agencies who noted how the research findings informed service or program modifications. Within SSA, there is now more focus on youth in the delivery of benefits counseling. At other agencies, particularly at DOL, there are service guides to support the general implementation and service delivery to youth with disabilities.

1. SSA Programs: WIPA programs' enhanced benefits counseling focuses specifically on youth and families

The YTD findings informed modifications to SSA information materials and prompted a focus of benefits counseling services on youth through the Work Incentives Planning and Assistance (WIPA) program. Prior to YTD, there was more limited information available to support youth in transition. Based on the YTD experience, SSA enhanced the information it provides in the *Red Book* (2020e) and annual notices by adding information about the age-18 redetermination, work supports, and non-SSA programs (e.g., VR) that offer service to youth. For WIPAs, SSA implemented changes that placed more emphasis on youth in service delivery and began convening quarterly calls for WIPA staff to discuss youth transition services. Additionally, the WIPA program manuals were updated to include several references to youth and lessons from YTD.[20]

Despite the changes prompted by YTD, the PROMISE experience suggests there remains a need to continue providing information and increasing access to benefits counseling for youth. When the PROMISE demonstration began, many WIPA programs contracted to provide youth services were not focused on youth receiving SSI. Benefits counseling was a critical service offered under PROMISE and one of the service types infrequently used by control group youth, contributing to the large impacts of PROMISE programs on families' use of this service.

---

[20] See, for example, *WIPA & Community Partner Work Incentives Counseling Training Manual,* Module 1. "Supporting Increased Employment and Financial Independence Outcomes for Social Security Disability Beneficiaries." Virginia Commonwealth University. Updated September 26, 2021. https://vcu-ntdc.org/resources/ntcmanual.cfm.

2. Other agencies: SSA demonstrations informed WIOA implementation and future policy proposals

Other agencies have used the findings from SSA demonstrations to influence their programs and policies. DOL cited lessons from YTD in working with other agencies to support the implementation of Workforce Innovation and Opportunity Act (WIOA) (DOL, "Employment," n.d.). Some of the PROMISE states incorporated features of their PROMISE interventions into their WIOA services. To continue policy development for youth, a DOL report (n.d.) cited the importance of work-based experiences and the continuity of supports through the transition period from YTD.

The PROMISE experiences have also influenced three policy proposals developed under DOL's SSI Youth Solutions project. Two of the proposals offer ideas for implementing and sustaining PROMISE-like services, particularly family case management, within the existing transition service landscape (Karhan and Golden 2021; Anderson, Hartman, and Ralston 2021). The third proposal describes a tool that links data across state and local programs to track the service delivery and outcomes of transition-age youth with disabilities (Gingerich and Crane 2021).

SUGGESTIONS FOR SSA'S FUTURE LEARNING AGENDA

We provide eight suggestions for SSA's future learning agenda to enhance outcomes for youth. The suggestions represent ways to address gaps in current knowledge and enhance the experiences of youth and their families in future demonstrations. We first discuss suggestions for future demonstrations and then offer ideas to test modifications to SSA program rules.

Future Demonstration Considerations

We identified four areas for future demonstrations to test interventions that support youth receiving SSI. They cover topics related to leveraging service models, data, and options to enhance outcomes that were not fully addressed in prior demonstrations.

Future Demonstration Considerations

1. Adapt existing models with strong evidence for SSI youth
2. Expand SSA data use among other public and private agencies
3. Identify and test interventions that improve family outcomes
4. Enhance understanding of diversity, equity, and inclusion

1. Adapt Existing Models with Strong Evidence for SSI Youth

A potentially useful starting point for future interventions is to adapt features of promising existing models to SSI youth. Year Up and Job Corps are examples of such models. SSA could identify and explore partnerships with entities implementing promising models to test whether referrals of SSI youth to those services and potentially augmenting the services with components, like benefits counseling, that

address needs that are unique to SSI youth, improves the outcomes of SSI youth. Programs that have not traditionally served youth with disabilities might need training or technical assistance to effectively serve youth who receive SSI. SSA's new Interventional Cooperative Agreement Program offers a possible vehicle for SSA to partner with non-federal entities to conduct tests of interventions for youth as well as other populations.

A related area for future learning might be to test the provision of longer-term supports to improve youth outcomes. All of the interventions we reviewed were of relatively short duration (less than five years). A successful transition to adulthood might require different types of services delivered at different stages of transition over a longer period. In later sections, we describe examples: a longer-term case management intervention that begins when families first enter the SSI program and a similar model delivered through an expanded WIPA role.

2. Expand SSA Data Use among Other Public and Private Agencies

One of SSA's key assets is its store of historical program and earnings data. These data allow SSA to track outcomes for decades after demonstration evaluations have concluded. For example, SSA has tracked the program and earnings outcomes of youth participating in YTD for ten years. However, the use of SSA data by other agencies to track the outcomes of populations they serve in common with SSA is more limited. This is partly due to the need to establish formal data use agreements, which can take many months, or years, to develop. Because of the complexity of SSA program data, their use also requires a substantial learning investment for other agencies to use them appropriately.

Linking SSA data with other programs can help SSA understand where SSI youth and families obtain services and how participation in those services affects youths' outcomes. An area for future consideration is how SSA might facilitate data-sharing agreements with other entities, building on its experience doing so for other research and demonstration efforts. For example, in PROMISE, SSA developed data use agreements with the 11 PROMISE states to link SSA data with state VR and Medicaid agency data. State and even private entities who have evaluation ideas could work more cooperatively with SSA to develop ways to share data to track how their interventions influence the long-term outcomes of youth receiving or at risk of receiving SSI. SSA has created data exchanges for operational purposes with several states (SSA, n.d.).[21] These exchanges are notable because they might already include

[21] Currently, federal, state, and local agencies interested in obtaining SSA data can request them by submitting SSA's data exchange request form (SSA-157). An outside agency's use of the data must be consistent with the administration of SSA's own programs, and the agency must meet SSA's data security requirements. There is also typically a cost in obtaining the data. The process for establishing a data exchange can take 18 months or longer to complete.

linkages across multiple systems to track cross program participation. Thus, one option would be for SSA to document and more widely publicize these linkages to encourage future usage. For example, if a state already has a data use agreement in place with SSA, it could amend it to incorporate its needs for delivering and evaluating transition services to youth receiving SSI.

With a more deliberate effort to link and share data SSA could be in a stronger position to support and learn from evaluations of efforts by public and private entities that affect SSI youth. SSA could also use the linked data in work with other agencies to track and report on cross-agency coordination efforts.

3. Identify and Test Interventions to Improve Family Outcomes

Another area for consideration relates to improving supports for family members. PROMISE emphasized family-focused case management and offered services to all family members of the youth receiving SSI. The early findings from PROMISE and the theoretical literature suggest that there is merit to engaging parents and providing them with training and resources to improve outcomes for their youth receiving SSI. The literature also suggests that starting at age 14 might be too late, given the importance of parental influence and behavior in shaping children's outcomes.

A promising avenue is to build on supports SSA already offers (e.g., *Red Book* and annual notices). Currently, there is one intervention in the field that is providing more intensive updates to parents at key points throughout the child's development about available resources and actions they should be taking concerning the youth (Deshpande and Dizon-Ross 2020).

A second option is to provide more guidance and oversight concerning parental training and control of resources, especially representative payees. SSA currently requires representative payees to complete an annual accountability report (Form SSA-623-OCR-SM), but parents of a child who receives SSI are exempt from this requirement. The form asks representative payees to describe how the SSI payments were spent on behalf of the recipient. A similar type of form could be used to remind parents and hold them accountable for specific actions that would be in the best interests of the youth at different ages. For example, having parents describe what they are doing to support the youth's functional development, school completion, and participation in work-based learning experiences. It would be labor intensive for SSA to monitor and follow up on the information, but making parents consider the options for their youth on a regular basis, coupled with information about resources, might serve to nudge parents to act in ways that are beneficial to the youth.

A third option is providing case management to support parents' ability to navigate the support system for their youth with disabilities and invest in the youth's development, not unlike the case management provided under PROMISE. However, this case management would begin when the family first begins to receive SSI and

continue in some form until the child turns age 18.[22] As in PROMISE, a final potential area for further learning is how to improve the education and income of SSI parents, again, starting at the time when families first enter SSI.[23]

4. Enhance Understanding of Diversity, Equity, and Inclusion

A fourth area for future learning is enhancing understanding of diversity, equity, and inclusion in efforts to improve employment outcomes for youth receiving SSI (Gary et al. 2019; National Disability Institute 2020). As documented above, youth receiving SSI are more likely to live in low-income families, be non-White, and have lower educational attainment relative to youth and young adults in the general population. There are other characteristics related to the youth's identity, such as sexual orientation and disability, that are not well understood even in descriptive statistics. None of the SSA demonstrations we reviewed examined whether the enrollment and service delivery differed by racial or ethnic subgroups. Nor was there a review of how the specific demographic characteristics of the service providers might have influenced enrollment, service delivery, or outcomes. Thus, an area for future learning is to develop a better understanding of racial and other differences in the outcomes of youth receiving SSI and the potential sources of any differences. Such a focus would be consistent with the President's Executive Order on Advancing Racial Equity and Support for Underserved Communities through the Federal Government (Biden 2021). Disparities in service participation and outcomes might arise for many reasons, including the ways in which organizations recruit and deliver services to culturally diverse populations. However, one must first identify whether such disparities exist before being able to identify and address their causes.

Modifications to SSA Programs and Services

We offer three suggestions to learn more about the effects of changes to SSA program rules and services. The first offers options to test program rules through waiver-only demonstrations. The

Modifications to SSA Programs and Services

1. Tests to support changes to SSI program rules
2. Test expanded benefit counseling services
3. Expand testing of informational outreach

[22] Case management is not a current function of SSA, but SSA could contract for case management services in ways similar to how it funds state Disability Determination Services, WIPA services, and services provided by Protection and Advocacy for Beneficiaries of Social Security programs.

[23] These types of interventions might have limited potential for a couple of reasons. First, the child SSI payment's primary purpose is to provide parents with more resources to care for a child with a disability, which might include spending more time investing in (or spending time with) the child and less time in the labor market. Second, it is unclear if marginal improvements in parents' education and income will translate into meaningful improvements for the youth in the long term; the effect is less direct than interventions that seek to improve parent's knowledge and actions that directly affect the youth receiving SSI.

other two examine the potential for more proactive outreach to recipients to better inform youth and families about program rules.

1. Tests to Support Changes to SSI Program Rules

In the past, SSA has implemented waivers to program rules in testing interventions for youth, but it has not conducted tests of rule changes on its own. "waiver-only" demonstrations could provide SSA and policymakers important insights into how (untested) administrative changes affect outcomes. For example, there is no rigorous evidence on the potential effects of several SSI eligibility provisions, such as the age-18 redetermination, CDRs (particularly fluctuations in the timing of CDRs), and earnings reporting, that could influence youth employment and program outcomes.

Examples of waivers to SSA rules that could serve as a waiver-only test include:

- **CDRs.** SSA's discretionary funding for CDRs changes over time, which has implications for the number of CDRs completed. The result is that the CDR schedule might be unclear to many families, which can create challenges for their long-term outcomes. SSA could test waiving CDRs until age 18 for a subset of youth and compare their experiences and outcomes to those who undergo CDRs on a regular schedule.
- **Delaying the age-18 redetermination.** The age-18 redetermination is not aligned with other federal agencies' definitions of the age at which a child is considered to be an adult and no longer eligible for child services. For example, youth are eligible for special education services under the Individuals with Disabilities Education Act until age 22. SSA could test the impact of delaying the adult redetermination until age 22. This change, which is one of the policy proposals under DOL's SSI Youth Solutions initiative (Larson and Geyer 2021), would align the age of children across federal agencies and allow SSI youth to continue receiving income support for a longer period while they develop their capacity to support themselves as adults.
- **Employment and earnings of youth and parents.** The SSI program has several provisions that allow recipients to retain more of their benefits as their earnings increase, however, their use is limited. One option is to test eliminating earnings reporting for the youth and/or parent. Such a test would provide information to SSA about how the program rules influence the labor outcomes of children and their caretakers and demonstrate their capacity to work in the absence of work disincentives created by the program rules.

SSA could also conduct tests of the policy proposals included in its budget request described earlier (SSA 2020g). The findings would provide important exploratory evidence that would allow policymakers to examine modifications to programmatic

rules or changes in administrative funding (e.g., CDRs) that could affect benefit durations and eventual outcomes.

2. Test Expanded SSA Benefits Counseling Services

A second area for exploration is reviewing whether youth can benefit from expanded services from the WIPA program. Currently, SSA has limited means to provide or facilitate services to youth directly, which is why we suggest a focus on the WIPA program; it offers an existing mechanism SSA could use to deliver direct services. For example, in previous youth demonstrations (e.g., YTD and PROMISE), SSA relied on other entities to deliver case management, employment promoting, and other services. A demonstration could test expanding the role of WIPA programs to provide more intensive case management and referrals for youth and families. WIPA counselors trained specifically to serve youth receiving SSI and families in a manner as PROMISE did might be a means for sustaining PROMISE-like services. WIPA programs could provide more proactive outreach and comprehensive counseling for youth and families than they do currently and act as central means for families to obtain information and connections to services.

Although SSA could test such an intervention under its demonstration authority, a change in the legislation authorizing the WIPA program would be required to permanently implement such a program. Specifically, the Ticket to Work and Work Incentives Improvement Act of 1999 limits the total amount of WIPA funding to $23 million. Hence, Congress would seemingly need to expand this funding to support an expanded role of WIPA programs.

3. Expand Testing of Informational Outreach

A final idea is for SSA to substantially expand information outreach tests to youth and their families. SSA sends several types of notices to recipients and their families about SSI, including information on redeterminations. In other contexts, SSA has tested the impact of outreach to those receiving SSI and SSDI. For example, Zhang et al. (2020) found that more periodic reminders of wage reporting increased the likelihood of wage reporting by adult SSI recipients. SSA could assess whether a more frequent distribution of its current annual notices to youth could better preparing them for their age-18 redetermination. SSA could also test whether shortening the notices, which are 20 pages long, improves youth participation in work incentives and preparation for their redetermination.

CONCLUSIONS

SSA's youth demonstrations have generated evidence that has informed practices and policies in serving youth receiving SSI. The early demonstrations (STETS and TETD) proved the feasibility and importance of competitive employment placements for individuals with significant disabilities and demonstrated them to be a viable

alternative to sheltered work settings. In providing this evidence, these early demonstrations set the stage for later cultural and policy changes that reflect the idea that those with significant disabilities can work in integrated, competitive jobs. The more recent demonstrations (YTD and PROMISE) represent the most extensive and rigorous tests ever conducted of interventions for youth with disabilities. They contributed additional rigorous evidence on the effectiveness of employment-focused, comprehensive service interventions to the literature and policy considerations regarding how to improve youth outcomes. Their findings influenced the WIOA implementation and will likely continue to affect youth programs and policies in the future.

Despite the influence of the demonstration findings on broader policy, the evidence has not led to changes in SSI program rules or services. The lack of reform might be because the interventions tested exhibited diminishing impacts over time, particularly on employment. Additionally, other agencies would need to be involved in the delivery of the intervention services, which further complicates scaling. From a programmatic perspective, SSA has tested waivers to program rules though not separate from intervention services. Hence, it is not possible to say how specific programmatic rule changes, such as excluding earnings from benefit calculations, would affect youth outcomes or program costs from prior demonstrations.

A focal point going forward is to improve the long-term outcomes for youth receiving SSI. This issue is important given that many youth receiving SSI tend to have limited incomes into adulthood. In general, the most promising interventions that generated long-term impacts for other youth populations, including Job Corps and Year-Up, offered youth new opportunities for services and development that were not readily available in the existing service environment. We also offered up ideas for additional consideration to develop, form, and evaluation new interventions and cross-agency collaborations. Finally, we identified options to test waivers to programmatic rules governing earnings, CDRs, and the age-18 redetermination in ways that could inform the youth's long-term development.

In summary, efforts to help youth who receive SSI achieve their full potential are worthwhile. The descriptive evidence indicates that youth likely will continue to experience challenges in navigating a fragmented system of supports without more substantive intervention and programmatic reform support. There are promising avenues for moving forward, with priority needed on options that can demonstrate long-term impacts into adulthood. Our suggested recommendations around intervention and evaluation offer options to meet the goal of improving long-term adult outcomes for youth receiving SSI.

Chapter 6

Comment

Lucie Schmidt
Williams College

Youth Supplemental Security Income (SSI) recipients often struggle with the transition to adulthood. This period is often also difficult for their families, particularly given that a large share of income in these families comes from the children's SSI benefits. Because of these issues, a number of demonstrations have been aimed at improving this transition and at generating better adult outcomes for this population. This chapter by Wittenburg and Livermore provides an excellent introduction to the issues associated with this important population of SSI recipients, as well as to the lessons learned from four demonstration projects targeted directly at teenagers and young adults receiving SSI benefits: Structured Training and Employment Transitional Services (STETS), Transitional Employment Training Demonstration (TETD), Youth Transition Demonstration (YTD), and the Promoting Readiness of Minors SSI (PROMISE) demonstration.

Two findings that emerge from the youth transition demonstrations seem particularly important. First, we know that the SSI program is complex. While there are a number of supports available to recipients, services vary across localities and are fragmented in their delivery. The demonstrations show that services and supports are important, and that more intensive services seem to have larger effects. Some of the most encouraging findings in this area come from the PROMISE demonstration, which had a direct focus on case management, and on helping connect families to services they needed. PROMISE improved on previous demonstrations by having state agencies lead the implementation of services and requiring collaboration across agencies, reducing the complexity faced by recipients and their families.

A second important lesson is that children on SSI have different service and support needs than adults. This point is critical, and calls to mind the discussion by Berkowitz and DeWitt (2013) in their history of the SSI program, where they note that benefits for children with disabilities were "slipped into the legislation" (34) without full discussion of the implications. As Wittenburg and Livermore (in "Youth Transition") clearly lay out, successful interventions for children on SSI require an understanding of child development processes (both cognitive and noncognitive) and of parental investments. Both child development and parental investments might be affected by benefit receipt and by program rules.

The youth SSI demonstrations take place in the broader context faced by low-income families in the United States. A growing body of research points to a number of hardships faced by these families. For example, evidence suggests that individuals facing conditions of scarcity suffer reduced cognitive load and functioning (Mani et al. 2013). Low-income families endure longer waiting times for necessary goods and

services (Holt and Vinopal 2021). These hardships interact with the challenges faced by parents of children with disabilities, as well as with the complexity and fragmentation of services provided through SSI.

The youth SSI demonstrations can also be viewed through the lens of the growing literature on administrative burden. As described by Herd and Moynihan (2018), administrative burden occurs when the design of public programs makes it more difficult for families to access resources. This administrative burden comes in the form of learning costs (costs incurred in learning about program rules, what services might be available, and how to access those services), compliance costs (the burden of following program rules and regulations), and psychological costs (stigma from program receipt and stress due to dealing with administrative processes) (Moynihan, Herd, and Harvey 2015). All of these costs are likely to be high for youth SSI recipients and their families, and compounded by physical and mental disabilities. Herd and Moynihan (2018) make the important point that while this burden might not be intentional, it is constructed by the way government programs are designed, and can therefore be reduced by policy interventions.

When we think about SSI children in the context of these broader challenges, two important questions arise: First, we know that the complexity and fragmentation of SSI-related services and supports generates administrative burden for youth recipients and their families. To what extent should SSI interventions be trying to reduce these burdens? And second, to what extent is the success of SSI youth demonstrations dependent on the ability to do so? Some of the most successful elements of the youth demonstrations actively seek to reduce administrative burden for youth SSI recipients and their families. In particular, the focus of the PROMISE demonstration on family case management and the push for collaboration across state agencies are likely to meaningfully reduce administrative burden and to improve outcomes for children and their families.

The introductory chapter of this volume looked at the body of evidence from the overall demonstrations and suggested that we ask big picture questions about the goals of these demonstrations. These kinds of questions could be particularly helpful in the youth SSI context. For example, what would a successful transition to adulthood for children on SSI look like? While several of the previous demonstrations focused on employment and earnings, the excellent discussion in this chapter about the specific needs of child SSI recipients suggests that perhaps a broader set of outcomes might be worth targeting. For example, Wittenburg and Livermore point out that children with disabilities might require additional parental investments at critical ages to support their development, but that the low-income parents of children on SSI might be constrained in making those additional investments. Should SSI demonstrations directly target development of cognitive and/or noncognitive skills, with the understanding that improving those skills is likely to improve adult outcomes? Should SSI demonstrations directly target parental investments? Quasi-experimental evidence suggests that SSI income for low birth weight infants can improve measured parenting

behavior (Guldi et al. 2018), but additional evidence from targeted demonstrations would be helpful.

Thinking about a broader set of outcomes also raises additional questions for future demonstrations. The demonstration evidence shows the importance of interventions that focus on both the SSI recipient and their family. But what is the optimal timing of interventions? If interventions are focused narrowly on preparing youth SSI recipients to enter the labor market, then it makes sense to begin them during the late teens. However, as Wittenburg and Livermore point out, starting in the teen years might be too late given the importance of the role of parents and of early investments in children. Some of the most exciting efforts mentioned in this chapter are interventions aimed at younger children and their families. For example, Deshpande and Dizon-Ross (2020) are providing additional information to families at important points during the child's development. Other possible interventions described would expand on the family case management found to be successful in the PROMISE demonstration, but instead begin when the child first begins SSI benefit receipt. Overall, the evidence from the youth SSI demonstrations suggests a number of ways in which future interventions can be used to improve the outcomes for children with disabilities and their families.

Chapter 6

Comment

Manasi Deshpande
University of Chicago

Wittenburg and Livermore (in "Youth Transition") provide a comprehensive overview of demonstration projects aimed at improving the outcomes of youth receiving Supplemental Security Income (SSI) benefits. Their chapter demonstrates that existing Social Security Administration (SSA) demonstrations focused on this population, including the Youth Transition Demonstration and the Promoting Readiness of Minors in SSI (PROMISE) demonstration, have provided important lessons on helping young people successfully transition to adulthood. The existing demonstrations have also highlighted critical areas of focus for future demonstrations. In this discussion, I put proposed future demonstrations, including those discussed in the chapter, into a broader conceptual framework and discuss how to prioritize the demonstrations.

DETERMINING THE GOALS OF THE SSI PROGRAM

The first step in prioritizing demonstration projects involving youth receiving SSI benefits is to determine the goals of the SSI program. As with most social safety net programs, policymakers may have several different objectives for the SSI program: providing income to recipients that is sufficient for their consumption and well-being, encouraging recipients who can work to work, and limiting program expenditures. These goals may be in conflict:

- SSI could provide a sufficient income and limit expenditures by, e.g., phasing out benefits quickly as recipients earn money in the labor market. However, this policy could discourage recipients from working.
- Alternatively, SSI could limit expenditures and encourage recipients to work by, e.g., cutting SSI benefits. However, this policy would provide less income to recipients.
- Or SSI could encourage recipients to work and provide sufficient income by, e.g., providing work subsidies and supports. However, this policy would likely increase program expenditures.

The priority for potential future demonstration projects depends on which goals are considered most important. This decision in turn requires evidence on the work capacity of youth receiving SSI benefits when they turn age 18. To this end, I present three potential models of SSI youth transition that speak to the work capacity of this population.

POTENTIAL MODELS OF SSI YOUTH TRANSITION

The three potential models of SSI youth transition tell different stories about the work capacity of youth receiving SSI benefits when they reach 18 years of age. Which model is correct has implications for the goals of the SSI program and therefore for demonstration project priorities. I discuss each model in turn and then discuss evidence on which model best reflects reality. Of course, it could be that different models apply to different parts of the SSI youth population. In that case, the goal would be to determine which model is most prevalent, or whether observable characteristics can predict which model is relevant for a particular child.

Model 1: The SSI Children's Program Is Well Targeted, So Youth Who Receive SSI Benefits Have No or Little Work Capacity at Age 18

Under this potential model, SSA excels at identifying and enrolling youth who are likely to have limited work capacity as adults, because of either disability or poverty or both. This would mean that youth who receive SSI benefits have no or little work capacity at age 18—not because of the effects of the SSI program, but simply because of selection. If this model is correct, then it could be reasonable for SSA to focus on the goal of providing recipients with a sufficient income. In this case, the most relevant demonstration projects would be those that keep more youth on SSI for a longer period, and those that phase benefits out quickly as earnings increase. Specific demonstrations include:

- Pushing the age-18 redetermination to age 22 or above
- Changing the age-18 redetermination criteria to weight vocational factors (such as skills) more heavily

Model 2: Youth Who Receive SSI Have Work Capacity at Age 18 but Avoid Productive Activities Like Work and School in Order to Demonstrate Disability and Stay on SSI

Under this potential model, youth who receive SSI have work capacity at age 18 but intentionally limit their productive activities such as school and work out of fear of losing their SSI benefits. If this model is correct, then it could be reasonable for SSI to focus on encouraging work among transition-age recipients. In this case, the most relevant demonstration projects would be those that reduce explicit and implicit work penalties or even subsidize work, and those that build a stronger safety net outside of SSI. Specific demonstrations include:

- Cut SSI benefits or turn a fraction of them into work supports
- Raise Substantial Gainful Activity (SGA) or change the way SGA capacity is assessed for SSI youth transitioning to the adult program[24]

Although building a stronger safety net outside of SSI is outside of SSA's scope, natural experiments could provide evidence on whether strengthening the safety net outside of SSI can encourage work. For example, the expanded child tax credit and recent Medicaid expansions could make losing SSI less consequential and thereby encourage work.[25]

Model 3: Youth Who Receive SSI Are Physically Capable of Work but Lack the Skills to Work

Under this potential model, youth who receive SSI are physically capable of work but lack the skills to be productive in the labor market. If this model is correct, then the logical goal of SSI is building skills early to encourage work later. One possible approach is to build on existing demonstration projects such as YTD and PROMISE that intervene in adolescence and conduct demonstration projects that intervene earlier (e.g., in early childhood). Specific demonstrations include:

- Eliminate child continuing disability reviews, which could potentially encourage skill formation
- Sponsor skill-building programs, such as literacy and "intensive" or "high-dosage" tutoring starting from a young age. The PROMISE demonstration finds that most youth receive several services (Mamun et al. 2019), so it would be important not to duplicate them[26]
- Provide information to families to create realistic expectations about whether children will receive SSI benefits as adults

---

[24] For example, regarding step 5 of the disability determination process, CFR §404.1566 states: "We will determine that you are not disabled if your residual functional capacity and vocational abilities make it possible for you to do work which exists in the national economy, but you remain unemployed because of—(1) Your inability to get work; (2) Lack of work in your local area; (3) The hiring practices of employers; (4) Technological changes in the industry in which you have worked; (5) Cyclical economic conditions; (6) No job openings for you; (7) You would not actually be hired to do work you could otherwise do; or (8) You do not wish to do a particular type of work." Some of these factors could be modified to take into account barriers to employment for youth receiving SSI benefits, such as labor market discrimination, difficulty of moving to another area, or inadequate skills or preparation for the labor market.

[25] Schmidt, Short-Sheppard, and Watson (2020) find no effect of ACA Medicaid expansions on disability applications.

[26] See, for example, Nickow, Oreopoulos, and Quan (2020).

WHICH MODEL OF YOUTH TRANSITION BEST REFLECTS REALITY?

As each of the three models has different implications for SSI goals and demonstration projects, it is important to determine which model best reflects reality. Though more research is needed to answer this question, current research finds substantial heterogeneity in the outcomes of youth receiving SSI benefits. YTD finds low baseline rates of employment and a minimal long-term impact of supports and services on employment rates. Similarly, Deshpande (2016a) finds that the vast majority of youth who are removed from SSI at age 18 do not earn anywhere close to SGA levels, even though they were removed because they were determined to be capable of SGA. For this group who are unlikely to earn at self-sufficiency levels even when not receiving SSI, Models (1) and (3) are most relevant. From Deshpande (2016a), about 20 percent of youth who are removed from SSI do earn at SGA levels in adulthood. For this group, Model (2) is likely the most relevant. However, it is difficult to predict using characteristics in SSA data which group a particular child receiving SSI will fall into. Improving this prediction exercise could improve the targeting of demonstrations and policies for youth receiving SSI benefits.

Chapter 6

Comment

Jennifer Sheehy
US Department of Labor[27]

The chapter authored by Livermore and Wittenburg ("Youth Transition") is an excellent summary of the state of the science on young Supplemental Security Income (SSI) recipients' transition to adulthood. As they note, youth with disabilities often have different needs from adults and, regardless of whether they receive SSI, they face a confusing set of services with different definitions, timelines, and rules. The authors highlight three strategies that can help young people with disabilities obtain employment as they transition into adulthood:

1. Place a strong emphasis on employment services, e.g., skill assessments, career aspirations, educational goals, on-the-job training, post-employment services and follow-ups.
2. Provide customized supports for youth, e.g., meet health care needs, reasonable accommodations, employment supports, and community and governmental transition supports.
3. Focus on providing services to the entire family unit, e.g., provide case management to youth and their family members; benefits counseling; financial education; career training for youth; parental training on available supports for their youth; and career resources for parents.

However, there are many challenges in implementing these strategies. Chief among them is that existing systems are fragmented, creating challenges for ensuring that youth have access to supports they need to be successful. For instance, some families have difficulty accessing all the services offered under past demonstrations. There may be administrative burdens, trust issues, different expectations about what will be offered or can be accomplished, child safety, and other concerns. An open question is how to structure programs and systems to ensure all families and youth have access to promising practices. It is important not to think of supports for youth with disabilities as standalone policies; these supports are often most effective when integrated throughout general youth-related policies.

Post-pandemic, America's recovery needs to be powered by inclusion. Focusing on improving services and programs for youth with disabilities is critical as we recover from the COVID-19 pandemic to avoid simply returning to the status quo, which failed many young people with disabilities. As a country, we have a unique opportunity to

[27] The views expressed in this chapter are those of the author and do not necessarily represent the views of the Department of Labor or the US federal government.

build upon what works to ensure systems are more inclusive and support all youth with disabilities.

The SSI Youth Solutions effort from the US Department of Labor (DOL), Office of Disability Employment Policy (ODEP), is developing knowledge by engaging subject matter experts to develop 12 novel policy, program, or service solutions to improve employment outcomes for youth with disabilities who apply for or receive SSI. The proposals are diverse—including training and apprenticeship transition supports, case management models, and postsecondary education and employment training curricula. DOL is currently assessing these proposals to determine their likely effectiveness and estimated cost for demonstration projects. These new projects may provide substantial opportunities for future SSI Youth demonstration efforts.

There has been a broad movement toward increased cooperation across agencies over the past decade or more, which presents new opportunities to increase coordination of services and reduce the thicket of fragmented programs faced by youth. DOL is committed to working with the Social Security Administration in the future to contribute to novel interagency efforts as part of the current administration's commitment to making our country more equitable and inclusive.

Chapter 7

An Overview of Current Results and New Methods for Estimating Heterogeneous Program Impacts

Till von Wachter
University of California Los Angeles
National Bureau of Economic Research

The numbers of beneficiaries of the Social Security Disability Insurance (SSDI) and recipients in the Supplemental Security Income (SSI) programs grew rapidly over the past decades.[1] At the same time, the demographic characteristics and impairment types of beneficiaries have evolved (Duggan and Imberman 2009; Duggan, Kearney, and Rennane 2015). These trends have raised the question whether beneficiaries and recipients might have greater potential to work now compared to in the past (e.g., Autor and Duggan 2006; von Wachter, Song, and Manchester 2011). To study how to best encourage and support employment among potentially able SSDI beneficiaries and SSI recipients, the Social Security Administration (SSA) has engaged in a series of demonstrations aimed at establishing the effect of various policy changes, incentives, and supports for SSDI beneficiaries' and SSI recipients' employment.

This chapter discusses to what extent the effect of the various interventions tested varies across subgroups of SSDI beneficiaries and SSI recipients. This is an important question because the SSDI and SSI programs insure and serve a broad population of individuals. Current beneficiaries and recipients not only vary substantially in their education, occupation, and skill backgrounds, but they also vary in age, gender, types of impairment, and time spent in the program. Trends that have raised the number of beneficiaries and recipients who are younger and/or are more likely to have impairments associated with musculoskeletal or mental health conditions have further increased that diversity. These are all factors that potentially affect their ability to work and to find work, as well as their likelihood of sustained success in the labor market.

SSA has pursued several demonstrations that aim to provide insights on a range of questions regarding key subgroups of the SSDI and SSI populations. How different beneficiaries and recipients respond to treatments tested in a demonstration is important for several reasons. Documenting the range of possible responses to treatments is helpful for better predicting the potential impact of a tested intervention

[1] This increase reversed in 2013 or 2014 and the subsequent decline in participation may have different causes and policy responses. However, the demonstrations reviewed in this chapter are largely a response to the increase in participation, and the lessons from them apply mainly to that situation.

if it was to be offered to the full population of beneficiaries nationwide. Insights on variation in the effects of treatment for certain groups can help in better implementing interventions by informing which beneficiaries and recipients might be particularly responsive to new features of a program and for which the intervention could be further improved.

Evaluation research often considers the nature of treatment effect heterogeneity, such as impacts for subgroups (e.g., see Brock, Weiss, and Bloom 2013; Bell and Peck 2016b; Rothstein and von Wachter 2017). In practice, however, it is often difficult to estimate subgroup effects because of insufficient statistical power, usually due to having smaller sample sizes for subgroups relative to an evaluation's full sample. This limitation means that the role that such differential estimates can play—for example, in better targeting new interventions to particular beneficiaries—is often also limited. A growing literature on estimating heterogeneous treatment effects implies lessons for the next generation of SSA demonstrations. A data-rich environment (such as with some of the SSA demonstrations) combined with analytical/methodological developments suggests some particularly promising opportunities.

This chapter begins with a review of current evidence on the employment potential of SSDI beneficiaries and SSI recipients, with particular focus on variation across subgroups. Then it summarizes evidence of subgroup impact variation among recent SSA demonstrations testing interventions aimed at raising labor force participation and self-sufficiency. For SSDI and concurrent SSDI beneficiaries and SSI recipients, the chapter discusses estimates from the Benefit Offset National Demonstration (BOND) and its predecessor, the Benefit Offset Pilot Demonstration (BOPD); the Mental Health Treatment Study (MHTS); Project NetWork; the Accelerated Benefits (AB) demonstration; and the Promoting Opportunity Demonstration (POD). For SSI recipients, the chapter discusses the Transitional Employment Training Demonstration (TETD) and the Structured Training and Employment Transitional Services (STETS) demonstration.[2] Next the chapter reviews some recent methodological literature on estimating heterogeneous impacts and suggests lessons for future demonstrations. The final section draws some broad conclusions.

BACKGROUND ON VARIATION IN EMPLOYMENT POTENTIAL

Substantial research has analyzed how employment and earnings vary among individuals. Individuals' ability, capacity, and desire to work is sometimes referred to as their employment or earnings or work "potential." A range of factors typically influences such potential. For example, employment potential relates to individuals' ability and desire to work, which is influenced by their health and disability, innate capacity, and preferences, education, work experience, training, family status, child

[2] See Chapter 6 in this volume for more detail on SSI demonstrations.

care, and transportation. The institutional environment—such as taxes and the availability and value of public assistance and transfers—is also a factor.

Employment potential is likely also to be directly affected by SSDI or SSI program design. This is because, according to program rules, an excess level of earnings over some point triggers gradual removal from the program; moreover, receipt of benefits could reduce individuals' need to work.[3] This makes an empirical analysis of employment potential among SSDI beneficiaries and SSI recipients particularly difficult, as it involves an assessment of an inherently unobservable outcome: *What would the employment of a participant be in the absence of program benefits?*

Past research has aimed to estimate the employment potential of non-working SSDI beneficiaries and SSI recipients. Such estimates can provide an indication as to which individuals might be most responsive to inducements to return to the labor force, and at what level one should expect their employment or earnings to be. Individuals with higher work potential are likely to face lower barriers to employment and could be more responsive to financial inducements to return to work. In addition, information on beneficiaries' and recipients' work potential—or factors correlated with greater work potential—could help predict differential responses to non-monetary inducements or other supports to return to work, as well. Responses to monetary or non-monetary inducements to work can also vary across beneficiaries for reasons other than their employment and earnings potential. As discussed later in this chapter, reasons could include variation in how they understand the program or variation in how the program is implemented across time and across space.

Bound (1989) concludes that employment potential of the average beneficiary is small. Considering major impairment groups, age, and gender, von Wachter, Song, and Manchester (2011) find important variation in the employment rates of rejected applicants (which is, admittedly, only partially useful for understanding the employment potential of beneficiaries). Further, they find younger workers and workers with impairments related to the musculoskeletal system or to mental health have higher employment rates than do older individuals or those with impairments of the respiratory or circulatory system. Based on analyzing employment of rejected applicants, Maestas, Mullen, and Strand (2013) and French and Song (2014) also find

---

[3] The effect of the presence of SSDI or SSI benefits on the labor supply of beneficiaries is sometimes referred to as the "disincentive" effect. This includes a "substitution" effect (that arises because individuals would lose earnings if they were to work more, and hence they work less) and an "income" effect (that arises because individuals would like to work less but cannot because of low income; this effect arises even if individuals were to keep their SSDI or SSI benefits if they work above the SGA level). Strictly speaking, only the substitution effect is considered a program distortion. Though there are a few studies trying to isolate the substitution and income effects, recent research points to an important role of the latter (e.g., Gelber, Moore, and Strand 2017). This distinction, not further discussed in this chapter, could be relevant in its own right for program changes aimed at increasing work potential.

variation in employment potential by age and impairment type. In addition, they indicate some variation in estimated employment potential by prior income, with high-income individuals exhibiting lower employment.

Hemmeter and Bailey (2016) find that in the years after exiting the program, SSDI beneficiaries whose benefits were terminated due to a medical review had a relatively high incidence of earnings (i.e., any employment in a given year), but low rates of consecutive years employed and very low earnings. I suggest this finding reflects an upper bound for labor market outcomes of SSDI beneficiaries because it is for beneficiaries who are perhaps better off health-wise than the average. Hemmeter and Bailey (2016) also find that earnings and employment of those exiting the program can vary substantially with age, impairment type, and time spent on the program. Although some of the characteristics analyzed are correlated (i.e., younger individuals are more likely to have shorter program duration), the results offer some insights regarding variation in employment potential for some groups.

For example, individuals terminated with less than two years in the program are the highest-earning group considered and have about $18,000 annual earnings, whereas those with six or more years in the program have about $11,000 annual earnings. This difference is unlikely explained solely by age. Even among the highest-earning group, only 50 percent of individuals studied had earnings in the five years after program exit, indicating that even those SSDI beneficiaries with employment potential can face substantial labor market barriers and be at risk of financial hardship absent benefits.

Although studied less extensively than SSDI, participation in the SSI program can also affect the future employment prospects of recipients. For example, Deshpande (2016b) finds that children removed from SSI due to age 18 redeterminations recover only one-third of lost SSI cash income, and those who stay off SSI earn only $4,400 on average per year in adulthood. Davies, Rupp, and Wittenburg's (2009) descriptive analysis of human capital development among youth receiving SSI illustrates the heterogeneity among this population and emphasizes the importance of coordination, both contemporaneous and longitudinal, of programs and interventions aimed at supporting these youth.

Overall, employment potential among SSDI beneficiaries and SSI recipients varies in predictable fashion, among others with age and impairment type. However, the groups studied in the literature are quite coarse and far from what would be needed to meaningfully identify specific groups of individuals that could or should be targeted for employment incentives or services. Moreover, the role of different personal characteristics is typically studied separately, but the intersection is likely to be particularly informative about an individual's employment potential. The amount of heterogeneity documented in the demonstration reports provides a sense in which the likely opportunities and needs are likely to substantially differ among beneficiaries. A young beneficiary with an impairment related to mental health will likely have different needs than will an older beneficiary with an impairment of the

musculoskeletal system. Their needs are likely to differ further by years of education, profession, and labor market experience.

Trends in characteristics of SSDI beneficiaries and SSI recipients have tended to further increase the diversity in the characteristics of individuals that SSA serves. For example, the number of SSDI beneficiaries who are younger and have impairments associated with musculoskeletal or mental health conditions has increased in the 1990s and early 2000s (Duggan and Imberman 2009). Between 1988 and 2013, the share of SSI recipients who are younger than age 64 increased by 20 percentage points. The share with intellectual and mental health disorders was 57 percent of the SSI caseload for the working-age population in 2013 (Duggan, Kearney, and Rennane 2015). Because beneficiaries and recipients who are younger and have intellectual and mental health impairments are typically found to have higher employment potential, these changes likely increased the overall employment potential among SSI and SSDI recipients.[4] Whether these trends will continue is a matter of ongoing analysis. Although some researchers have warned the demographic trends may lead to unsustainable increases in SSDI caseloads over the long term (e.g., Autor and Duggan 2006), others suggest that these changes could be temporary, related to the aging of the baby boom generation (Congressional Budget Office 2012; Board of Trustees 2014) and to the increasing share of women in the labor force (Goss 2013). In either case, ongoing changes in the population and in the labor market—such as those brought by the COVID-19 pandemic—and in the SSDI and SSI programs will likely continue to affect the distribution of beneficiaries and recipients and with it the variation in employment potential.

DISCUSSION OF HETEROGENEITY IN ESTIMATES FOR DEMONSTRATION OUTCOMES

SSDI provides income to insured individuals who are unable to engage in substantial gainful activity (SGA) due to a medically determinable physical or mental impairment. SGA occurs if earnings exceed a monthly threshold (SSA 2020e).[5] The SSI program provides income to disabled individuals with limited economic resources, regardless of whether they qualify for SSDI based on their work history, or any individual age 65 and older with limited economic resources.

In addition to providing income support, SSA's programs aim to support the efforts of SSDI beneficiaries and SSI recipients of working age who desire to return

---

[4] However, Howard Goldman (Comment in this volume) in his helpful discussion raises the important point that for impairments related to mental health, the ability to sustain employment may vary over time as mental health conditions wax and wane.

[5] Individuals are eligible to receive SSDI benefits ("insured") if they have sufficient quarters of covered earnings (e.g., see https://www.ssa.gov/pubs/EN-05-10072.pdf). For employees, SGA means if working and making more than $1,310 per month in 2021 (or $2,190 for beneficiaries who are legally blind). For self-employed individuals, any month during which work exceeds 80 hours is considered a Trial Work Period month.

to work. SSA's programs do this through a number of work incentives policies and complementary programs providing counseling, among other services and supports. For example, the Trial Work Period allows SSDI beneficiaries to have a total of nine months in which earnings can exceed the SGA level (not necessarily consecutively) over a five-year period. If a worker completes the TWP, then they begin the Extended Period of Eligibility, which is a 36-month period during which the beneficiary is eligible to receive SSDI benefits if earnings drop below the SGA level in a given month.[6] These and other policies are designed to support beneficiaries' return to the labor force (SSA 2020).

Within this broad framework, SSA has implemented demonstrations to test how it might further support SSDI beneficiaries' and SSI recipients' return to work. The interventions studied in these demonstrations test a range of employment inducements and supports. These include, among others, monetary incentives to work above the SGA limit (as in BOND, BOPD, and POD), training (as in STETS), case management (as in MHTS and Project NetWork), job search assistance (as in STETS and TETD), and access to health care (AB). In several cases, demonstrations combine multiple treatments. The two subsections that follow provide a brief overview of these demonstrations and discuss relevant findings from the analysis of subgroups conducted for SSDI beneficiaries and SSI recipients, respectively. (The Appendix in this volume provides additional information about all of SSA's demonstrations.)[7] The last subsection summarizes some broader lessons and practical insights from this discussion.

General Considerations for Comparing Results across Demonstrations

The eight randomized evaluations of SSA's demonstrations considered here all to some degree addressed potential differences in the impact of the evaluated intervention across groups of individuals. The dimensions of heterogeneity varied across studies. Exhibit 7.1 indicates the groups that each of the demonstrations analyzed. In terms of demographic differences, seven out of eight studies differentiated by age, four by education, four by gender, and two by race/ethnicity. Seven out of eight studies differentiated among types of health impairments, five among types of benefit receipt

---

[6] Following the Extended Period of Eligibility (or its reentitlement period), if SSDI payments have stopped because a beneficiary's income is substantial, SSA gives them five years during which their benefits can be reinstated if they again stop working because of their disability. During the five-year period, SSA will not require them to file a new disability application to get benefits; this is called Expedited Reinstatement. For those workers who lost their entitlement to benefits but need to quit working for the same or related medical impairment, Expedited Reinstatement allows benefits to start again without their needing to submit a new application. See, for example, https://choosework.ssa.gov/library/fact-sheet-trial-work-period-twp or the *Red Book* (SSA 2020e; https://www.ssa.gov/redbook).

[7] Additional analysis of the subgroups, subgroup impacts, and differential subgroup impacts discussed in this chapter is available on request from the author.

(i.e., SSDI-only and SSDI/SSI concurrent enrollments), and five between having prior employment or not. To aid the exposition, the chapter will refer to the higher-ordered group as "category" (e.g., gender, impairment type) and the defining characteristic within each category as "subgroup" (e.g., women, musculoskeletal system impairments).

Exhibit 7.1. Subgroups Included in the Analysis, by Demonstration

	Age	Gender	Education	Employment	Impairment	SSI or SSDI Receipt	Other
AB	X				X	X	X
BOND (Stage 1)	X			X	X	X	X
BOND (Stage 2)	X		X	X	X	X	X
BOPD	X	X	X	X			X
MHTS	X	X	X		X		
Project NetWork					X	X	
POD	X		X	X	X	X	X
STETS	X	X	X	X	X	X	X
TETD	X	X	X	X	X	X	X

Note: Depending on the demonstration, "Other" includes Medicaid use, health (self-reported), body mass index, race/ethnicity, location, living arrangements, future expectations.

When comparing differences in the effect of inducements and supports of employment among beneficiaries between studies, one potential difficulty is that each study has slightly different definitions of its main outcomes. The most consistently available outcome across the studies I review is the total amount of earnings and the total amount of SSDI benefits paid, corresponding to the focus of the demonstrations on testing policies aimed at reintegrating beneficiaries into the labor market. Other frequently examined outcomes were employment and incidence of earnings above the SGA level.

To achieve a minimum amount of comparability among studies, the discussion focuses on those two outcomes common across almost all studies: total earnings and total amount of SSDI benefits paid. Yet there are still some differences in how these outcomes are defined. One observation for SSA to consider: whether future demonstrations should have greater consistency of outcome measures used.

Exhibit 7.2 (beginning on page 279) summarizes subgroup impacts from among the SSDI-focused demonstrations, and Exhibit 7.3 (beginning on page 293) summarizes subgroup impacts from among the SSI-focused demonstrations. In each exhibit, Panel A reports earnings impacts and Panel B reports benefits impacts. It is important to note that only the BOND report provided standard errors for the difference in the estimated impacts within a category, information needed to report whether the difference in the impacts *between* two subgroups is statistically significant. In the remaining cases, we can only assess whether the finding *within* a particular subgroup is statistically significantly different from zero, but not whether

there are statistically detectable differences *between* subgroups' impacts. This can be an important drawback for understanding subgroup heterogeneity and is further discussed in the last subsection ("Potential Insights and Practical Considerations").

BOND, MHTS, STETS, and TETD provide extensive subgroup impact estimates; whereas AB and Project NetWork provide mainly a description of subgroup-related findings. BOPD had a more limited exploration of subgroups. As result, the discussion will focus on BOND, MHTS, STETS, and TETD, with shorter mention of the other demonstrations. Preliminary results from POD's Interim Evaluation Report are discussed briefly, as well.

Demonstrations Focused on SSDI and Concurrent Beneficiaries

Benefit Offset National Demonstration (BOND)[8]

Informed by the results of its pilot study (BOPD), BOND tested the impact of a benefit offset on a nationally representative sample of SSDI (or concurrent SSDI/SSI) beneficiaries. In a first stage, the evaluation randomly assigned all SSDI beneficiaries in 10 randomly chosen SSA areas to either a treatment group that receives the offset and work incentives counseling (WIC) or a control group that receives only WIC. A second stage tested the impact of the offset on a group of SSDI-only beneficiaries who volunteered for the demonstration and thus were expected to be more likely to use the offset. In addition, a second stage tested the extent to which enhanced work counseling improves outcomes compared to WIC services by randomly assigning volunteers to either treatment 1 (benefit offset and WIC), treatment 2 (benefit offset and enhanced WIC), or a control group. All treated participants had access to the benefit offset during a 60-month participation period after completing the TWP.

[8] Discussion based on Gubits et al. (2018a/b).

Exhibit 7.2. Summary of Subgroup Analysis of Demonstrations Focused on SSDI Recipients

Panel A: Earnings						
Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration	Gender
BOND (Stage 1): Total Earnings						
Age ≤49	N/A	Employed	Affective disorder	SSDI	Short	N/A
T1: $9,328		T1: $28,954	T1: $7,336	T1: $7,276	T1: $7,687	
C1: $9,336		C1: $29,295	C1: $7,127	C1: $7,245	C1: $7,754	
Impact: –$8		Impact: –$341	Impact: $209	Impact: $31	Impact: –$67	
Age >49		Not employed	Other impairment[a]	Concurrent	Long	
T1: $4,212		T1: $2,026	T1: $6,499	T1: $3,724	T1: $6,175	
C1: $4,188		C1: $1,946	C1: $6,528	C1: $3,812	C1: $6,133	
Impact: $24		Impact: $80	Impact: –$29	Impact: –$88	Impact: $42	
Diff: –$32		Diff: –$421	Diff: $238	Diff: $119	Diff: –$109	
			Back disorder			
			T1: $4,819			
			C1: $4,724			
			Impact: $95			
			Other impairment[b]			
			T1: $6,929			
			C1: $6,933			
			Impact: –$5			
			Diff: $100			

Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration	Gender
BOND (Stage 2): Total Earnings						
Age ≤49 T21 & T22: $22,661 C2: $20,690 Impact: $1,971* Age >49 T21 & T22: $13,464 C2: $12,355 Impact: $1,109 Diff: $862	<Associate's T21 & T22: $15,746 C2: $13,797 Impact: $1,948* Any postsecondary degree T21 & T22: $22,724 C2: $21,665 Impact: $1,059 Diff: $889	Employed T21 & T22: $42,374 C2: $39,870 Impact: $2,504 Not employed T21 & T22: $9,618 C2: $8,236 Impact: $1,382 Diff: $1,123[c]	Affective disorder T22 + T21: $20,037 C1: $18,079 Impact: $1,958 Other impairment[a] T22 + T21: $17,573 C1: $16,136 Impact: $1,437* Diff: $521 Back disorder T22 + T21: $13,378 C2: $13,849 Impact: –$471 Other impairment[b] T22 + T21: $18,743 C2: $16,904 Impact: $1,839* Diff: –$2,310	SSDI T1: $19,522 C1: $17,348 Impact: $2,174 Concurrent T1: $17,109 C1: $15,901 Impact: $1,207 Diff: $966	Short T21 + T22: $18,968 C2: $17,348 Impact: $1,620* Long T21 + T22: $17,383 C2: $15,901 Impact: $1,482 Diff: $138	N/A

Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration	Gender
MHTS: Past 3 Month's Earnings (at study exit)						
Age 18-34 T: $913 C: $879 Impact: n.r. Age >35 T: $854 C: $443 Impact: n.r.* Diff: n.r.	<High school T: $747 C: $378 Impact: n.r.* High school T: $754 C: $328 Impact: n.r.* >High school T: $923 C: $568 Impact: n.r.* Diff: n.r.	N/A	Affective disorder T: $958 C: $500 Impact: n.r.* Schizophrenia T: $639 C: $421 Impact: n.r.* Diff: n.r.	SSDI only T: $2,374 C: $2,048 Impact: $326 SSI only T: $1,060 C: $893 Impact: $167 Concurrent T: $1,249 C: $1,279 Impact: –$30 Diff: n.r.	N/A	Men T: $862 C: $402 Impact: n.r.* Women T: $855 C: $550 Impact: n.r.* Diff: n.r.
POD: Earnings in 2019						
Age <50 T: $6,059 C: $6,131 Impact: –$71 Age ≥50 T: $3,711 C: $3,728 Impact: –$17 Diff: n.s.	≤High school T: $4,187 C: $4,209 Impact: -$22 >High school T: $5,929 C: $6,048 Impact: –$119 Diff: n.s.	Employed T: $13,416 C: $13,914 Impact: –$499 Not employed T: $2,217 C: $2,063 Impact: $154 Diff: n.s.	Mental T: $4,817 C: $4,752 Impact: $64 Musculoskeletal T: $4,398 C: $4,366 Impact: $32 Other[d] T: $5,138 C: $5,340 Impact: –$202 Diff: n.s.	N/A	N/A	N/A

Panel B: SSDI Benefits					
Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration
BOND (Stage 1): Total Benefits Due					
Age ≤49	N/A	Employed	Affective disorder	SSDI	Short
T1: $46,794		T1: $57,704	T1: $53,748	T1: $60,066	T1: $58,292
C1: $45,845		C1: $55,146	C1: $53,207	C1: $59,422	C1: $57,829
Impact: $949*		Impact: $2,558*	Impact: $541*	Impact: $644*	Impact: $463*
Age >49		Not employed	Other impairment[a]	Concurrent	Long
T1: $60,758		T1: $53,423	T1: $54,231	T1: $27,511	T1: $52,368
C1: $60,398		C1: $53,162	C1: $53,576	C1: $26,903	C1: $51,655
Impact: $359*		Impact: $261*	Impact: $655*	Impact: $608*	Impact: $714*
Diff: $590†		Diff: $2,297†	Diff: –$114	Diff: $37	Diff: –$251
			Back disorder		
			T1: $64,252		
			C1: $63,790		
			Impact: $462*		
			Other impairment[b]		
			T1: $52,496		
			C1: $51,829		
			Impact: $667*		
			Diff: –$205		

Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration
BOND (Stage 2): Total Benefits Due					
Age ≤49	<Associate's	Employed	Affective disorder	SSDI	Short
T21 & T22: $46,543	T21&T22: $48,664	T21 & T22: $47,863	T22 + T21: $49,292	T1: $53,965	T21 & T22: $53,958
C2: $44,373	C2: $47,049	C2: $43,973	C2: $48,210	C1: $52,361	C2: $52,361
Impact: $2,170*	Impact: $1,615*	Impact: $3,890*	Impact: $1,083	Impact: $1,604*	Impact: $1,597*
Age >49	Any postsecondary degree	Not employed	Other impairment[a]	Concurrent	Long
T21 & T22: $56,434	T21 & T22: $57,421	T21 & T22: $52,825	T22 + T21: $52,033	T1: $49,928	T21 & T22: $49,724
C2: $54,861	C2: $55,116	C2: $51,644	C2: $49,971	C1: $47,652	C2: $47,652
Impact: $1,574*	Impact: $2,305*	Impact: $1,180*	Impact: $2,062*	Impact: $2,277*	Impact: $2,072*
Diff: $597	Diff: –$690	Diff: $2,709†	Diff: –$979	Diff: –$672	Diff: –$476
			Back disorder		
			T22 + T21: $54,175		
			C2: $53,329		
			Impact: $846		
			Other impairment[b]		
			T22 + T21: $51,096		
			C2: $49,085		
			Impact: $2,011*		
			Diff: –$1,165		

Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration
MHTS: Past Month's SSDI Benefit Amount					
Age 18-34	<High school	N/A	Affective disorder	SSDI only	N/A
T: $689	T: $748		T: $896	T: $630	
C: $680	C: $701		C: $887	C: $628	
Impact: n.r.	Impact: n.r.		Impact: n.r.	Impact: $2	
Age ≥35	High school		Schizophrenia	SSI only	
T: $871	T: $800		T: $767	T: $36	
C: $868	C: $790		C: $759	C: $42	
Impact: n.r.	Impact: n.r.		Impact: n.r.	Impact: –$6	
Diff: n.r.	>High school		Diff: n.r.	Concurrent	
	T: $899			T: $378	
	C: $913			C: $383	
	Impact: n.r.			Impact: –$5	
	Diff: n.r.			Diff: n.r.	

Age	Education	Employment Status	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	SSDI Benefit Duration
POD: SSDI Benefit Amount (in 12 months after enrollment)					
Age <50	>High school	Employed	Mental	N/A	N/A
T: $11,037	T: $13,248	T: $10,641	T: $11,229		
C: $10,997	C: $13,447	C: $10,850	C: $11,125		
Impact: $40	Impact: –$199	Impact: –$209	Impact: $103		
Age ≥50	≤High school	Not employed	Musculoskeletal		
T: $12,890	T: $11,214	T: $12,401	T: $12,738		
C: $12,956	C: $11,092	C: $12,352	C: $12,656		
Impact: –$66	Impact: $122	Impact: $49	Impact: $82		
Diff: n.s.	Diff: n.s.	Diff: n.s.	Other impairment[d]		
			T: $12,338		
			C: $12,489		
			Impact: –$152		
			Diff: n.s.		

Notes: Due to rounding, values that were computed at greater levels of precision may appear not to sum as whole numbers.
The "Impact" is the treatment-control difference; the "Diff" is the between-group difference in subgroup impacts.
n.r. indicates that the evaluation did not report the value (but in some cases, it did report whether the value was statistically significantly different from zero and is noted).
n.s. indicates that the diff value, not reported, is not statistically significantly different from zero.
Statistical significance is indicated (per the specific evaluation's conventions) as follows: * = impact is statistically significantly different from zero. † = diff (impact differential between subgroups) is statistically significantly different from zero.
[a] "Other impairments" denotes everything other than affective disorder.
[b] "Other impairments" denotes everything other than back disorder.
[c] *p*-value was 10.3%.
[d] "Other impairments" denotes everything other than mental disorder or musculoskeletal disorder.

BOND was notable due to its large sample sizes. The primary outcomes of the study were a cumulative earnings measure (2011–2015 for Stage 1 and 2012 and 2015 for Stage 2) and total SSDI benefits (as recorded in May 2017).[9] BOND did not find any evidence of the benefit offset policy increasing total earnings or decreasing total SSDI benefits in either stage. In fact, the evaluation found strong evidence of the offset policy increasing SSDI benefits in both stages.[10] The different forms of work incentives counseling in Stage 2 treatments did not have any effect on earnings or SSDI benefits.

BOND is by far the largest experimental demonstration reviewed here, and therefore its analysis would be expected to be most likely to detect differences across groups. As noted earlier in Exhibit 7.1, Stage 1 of the demonstration evaluated the effect of treatment across the following subgroups: age, employment status, type of health impairment, SSDI benefit duration, SSI status, and access to a Medicaid buy-in program. Stage 2 of the demonstration evaluated impacts for a slightly different set of subgroups: age, employment status, type of health impairment, SSDI benefit duration, and access to a Medicaid buy-in program.

Looking across estimates in Exhibit 7.2 of the effect of the benefit offset on earnings from the BOND study, it appears that overall there are no cases in which impact estimates are found to be statistically different within categories. The only *difference* in subgroup effects within categories that approaches significance relates to impact estimates by prior employment in Stage 2. Those who were not employed at baseline had an earnings impact, whereas those who were employed at baseline did not. The difference between these impacts approaches statistical significance (with a *p*-value of .103). Next, in Stage 2 of BOND, it appears that impact estimates for younger beneficiaries (age 49 and younger), beneficiaries with less than an associate's degree, beneficiaries in the subgroup with a primary impairment other than a major affective disorder, and beneficiaries in the subgroup with a primary impairment other than a back disorder were found to be positive and statistically significantly different from zero. (No subgroup estimates were statistically significantly different from zero in Stage 1 of BOND.)

Although not statistically significant, the differences in impact estimates between some of the other subgroups are substantial—for example, younger beneficiaries have

---

[9] Previous BOND reports used a different outcome measure, SSDI benefits paid. SSA occasionally makes incorrect payments to beneficiaries and later corrects for these payments. "Benefits paid" is the value SSA paid a beneficiary at the time; "benefits due" is a revised measure of the amount a beneficiary should have received at the time.

[10] This can be explained by the combination of a larger positive mechanical effect and a smaller negative behavioral effect on SSDI benefits. The second of these effects (which implies individuals moving from full benefits to partial benefits because of increased employment) is swamped by a larger amount of individuals already working at the SGA level who mechanically go from zero benefits (due to suspense under current-law rules) to partial benefits (because of the benefit offset). Given the structure of the offset, there needed to be more beneficiaries moving into SGA from nonemployment for average benefits to go down.

nearly double the earnings increase of older beneficiaries, and the same is true for more- versus less-educated beneficiaries. These are important findings, as they show that in contrast to the zero average effect, certain salient subgroups appear to have experienced increases in earnings in response to the benefit offset.

Considering effects on receipt of SSDI benefits in Panel B of Exhibit 7.2, almost all subgroup-specific estimates are statistically significant in both stages of BOND. These effects are all positive, indicating that the benefit offset raised rather than lowered SSDI benefit amounts.[11] In three instances, the differences within categories are statistically significant. For example, in both stages, we can see that beneficiaries with prior employment (compared to those without) have substantially larger increases in SSDI benefits (both in absolute and percentage terms). Similarly, there is a difference between impact estimates for older and younger people in Stage 1, but not Stage 2.

Interestingly, for both the employment status and the age group comparisons, the groups with higher earnings impacts also have higher impacts on SSDI benefits received. In contrast, for education groups or SSDI-only versus concurrent beneficiaries, those groups with higher earnings impacts have lower (albeit still positive) impacts on SSDI benefits.

The BOND study also evaluated subgroup impacts for other outcomes. In total, 364 tests of difference in impacts were conducted for Stage 2 subgroup analysis, implying that some of the tests would be statistically significant by chance. Yet, there was no clear pattern of the offset's behavioral effects in the subgroup analysis beyond those already discussed. Weak evidence is presented that the effect on employment and earnings above the SGA level outcomes is greater for participants with less education (statistically significant in 2 out of 12 tests; see Gubits et al. [2018b, Exh. F-49, F-50, and F-52]). This subgroup had lower rates of employment overall, which the report suggests could have led to larger effects for this group.

Mental Health Treatment Study (MHTS)[12]

The MHTS tested how access to supported employment services and systematic medication management services affects the ability of SSDI beneficiaries with schizophrenia or an affective disorder to return to work. The treatment group received a comprehensive package of mental health and employment services and was exempted from medical continuing disability reviews for a three-year period after study enrollment. The control group was given a list of available local and national resources along with a $100 payment for participating in quarterly interviews and was not exempted from medical continuing disability reviews. Relative to BOND or Project NetWork, MHTS was a relatively small evaluation, involving 2,238 volunteers

[11] This is consistent with the intent of BOND to leave participants on average better off than nonparticipants by allowing them to keep receiving some benefits while working.

[12] Discussion based on Frey et al. (2011).

among SSDI beneficiaries between the ages of 18 and 55 diagnosed with either schizophrenia or an affective disorder from 23 study sites. Once enrolled, participants remained in the study for two years.

Given MHTS's focus on employment and health outcomes for SSDI beneficiaries, the primary outcomes of interest were a participant's monthly employment rate, self-reported physical and mental health scores, and quality of life. The study had a number of other exploratory outcomes related to employment and health. MHTS's intervention had substantial positive effects on the employment rate and earnings but did not lead to a statistically significant reduction in SSDI benefits. The study found that the mental health score (but not the physical health score) and general life satisfaction improved for the treatment group relative to the control group during the study period.

MHTS is another demonstration that reports detailed subgroup impacts. As shown earlier in Exhibit 7.1, the demonstration evaluated treatment impacts for the following subgroups: age, gender, educational attainment, and disorder diagnosis. The report did not provide test statistics that would allow us to assess whether impact estimates across subgroups within categories were statistically different from one another. Comparing between earnings (Panel A) and SSDI receipt (Panel B), for MHTS the opposite pattern from BOND emerges. There are no detectable subgroup impacts for SSDI receipt; however, there are several instances of subgroup impacts for earnings.

MHTS distinguishes between earnings averages that consider the whole sample (i.e., include zeros for those who are not employed; called "unconditional" estimates) and averages that consider those who are employed (i.e., exclude zeros for those who are not employed; called "conditional" estimates). Because the observed impacts on employment imply the estimates of the conditional earnings could be based on a selected sample of workers, this chapter reports just the unconditional (experimentally valid) impacts.

These earnings impacts appear to be larger for older workers than for younger workers, the latter being the only group that does not have a detectable increase. All three of the education groups considered experienced earnings impacts. These earnings impacts are positive (and precisely estimated) also for those with affective disorder and schizophrenia, with the former group experiencing somewhat larger increases.

Overall, these impact estimates suggest the treatment was broadly successful in raising employment for the population of eligible beneficiaries, with young beneficiaries being a clear exception. As confirmed by Panel B, in none of the subgroups did the rise in earnings lead to a reduction in SSDI benefits.

Other SSDI-Related Demonstrations

In contrast to BOND and MHTS, BOPD, and Project NetWork did not engage in a systematic analysis of subgroup impacts. The remainder of the section summarizes some of the results from these three evaluations as discussed in the respective reports. Preliminary results from POD are mentioned, as well.

Benefit Offset Pilot Demonstration (BOPD).[13] The four-state BOPD was the pilot study for BOND. BOND tested the effect of an intervention that reduced annual SSDI benefits by $1 for every $2 of annual earnings above an annualized measure of SGA. By providing beneficiaries with a "ramp" that gradually reduces SSDI benefits as earnings increase, the benefit offset treatment prevented SSDI beneficiaries (or concurrent SSDI/SSI beneficiaries) earning more than the SGA amount from facing a "cash cliff" after the TWP runs out. From 2004 to 2010, BOPD randomly assigned a group of volunteers from four states (CT, UT, VT, WI) to either a treatment group (N=917) or a control group (N=893). The benefit offset was available to volunteers assigned to the treatment group for a six-year period after they completed the nine-month TWP. The pilot focused on working through issues in administering the offset, but it also measured the impact of the offset on average annual earnings, the incidence of having any employment in a given year, and annual SSDI benefits received by volunteers as measured by administrative records.

The benefit offset increased the proportion of beneficiaries with earnings above the SGA level in the two years after random assignment (Weathers and Hemmeter 2011). However, the benefit offset did not result in reductions in SSDI benefit payments. Moreover, among beneficiaries who made more than the SGA level before random assignment, the benefit offset provisions reduced average earnings.

Each state also conducted its own evaluation, including subgroup analyses. For example, in Wisconsin, subgroup analysis was conducted for six pairs, including analysis of age, gender, earnings history (any earnings and $1,200 cutoff), Medicaid buy-in,[14] and TWP completion (Delin et al. 2010). In terms of annual earnings, out of 108 subgroups, the 5 for which there was an impact were women, the "no Medicaid buy-in" group, those with "pre-enrollment earnings," and those with "less than $1,200 pre-enrollment earnings." All of these impacts occurred in the first and second quarter after study enrollment, with no detectable impacts from the third quarter to the eighth quarter.

Connecticut and Vermont exhibited similar trends of early impacts for the Medicaid buy-in subgroup, and Connecticut also reported impacts in the older subgroup (Porter et al. 2009; State of Connecticut 2009). Although there were no detectable impacts for the Medicaid buy-in or age subgroups in Utah, there were impacts for men and subgroups based on pre-program earnings (Chambless et al. 2009).

The subgroup analysis did not examine impacts on SSDI benefits. With the exception of the Medicaid buy-in distinction, the subgroup analyses do not change the impression from BOND, which was a larger experiment that tested similar subgroups. BOPD's subgroup analyses do reflect, however, the presence of differences across treatment sites also seen in other demonstrations.

---

[13] Discussion based on Weathers and Hemmeter (2011).

[14] Medicaid buy-in programs allow workers with disabilities access to Medicaid services.

Project NetWork.[15] The Project NetWork demonstration studied the impact that case management had on earnings and SSDI receipt for beneficiaries with severe disabilities. Implemented in 1991 at eight sites across the United States, Project NetWork recruited 8,248 volunteers of eligible SSDI (or concurrent SSDI/SSI) beneficiaries. Participants were randomly assigned into the treatment group or the control group. The demonstration provided treatment group members with three services: outreach, waivers, and case management. The study used four different case management models. All models had the same outreach procedures and work incentive waiver provisions but differed in the implementation of case management. Each model was implemented in two of the eight sites.

Project NetWork measured the impact of treatment on average monthly disability benefits, receipt of services, and average earnings two years after study enrollment. The program overall achieved annual earnings gains in the first two follow-up years (but not in the third). Project NetWork did not reduce the amount of SSDI or SSI receipt.

The Project NetWork evaluation analyzed impacts for subsets of the sample defined by type of eligibility (SSDI-only versus SSI-only versus concurrent) and primary impairment (mental versus neurological versus musculoskeletal versus other). Average annual earnings increased for SSDI-only beneficiaries. SSI-only recipients, and concurrent beneficiaries did not have a detectable earnings impact, which could reflect prior work experience or imply that SSDI-only beneficiaries required fewer services in order to return to work.

Important to this chapter is that no test for the difference between groups' impact estimates was provided. Discussion of differences in impact estimates by impairment type noted that treatment generally did not affect earnings or SSDI benefits for any of the impairment types considered (Kornfeld and Rupp 2000).

Accelerated Benefits (AB).[16] The AB demonstration tests how early access to medical services affects new SSDI beneficiaries' health and employment outcomes. The demonstration provided new SSDI beneficiaries with access to AB health care during the 24-month waiting period before their transition to Medicare. The demonstration randomly assigned 1,997 participants into three groups: in addition to standard SSDI benefits, treatment group 1 (AB) received access to AB health care whereas treatment group 2 (AB Plus) received AB health care plus telephone counseling services; the control group received only SSDI benefits.

The primary outcomes focused on three health factors: health care use, unmet need, and health status. The exploratory variables focused on employment. The main results for the primary outcomes of the demonstration were a rise in health care utilization and a reduction in the share of participants with unmet medical needs. Although the AB Plus group tended to look more for work and used more Ticket to

[15] Discussion based on Kornfeld and Rupp (2000).

[16] Discussion based on Michalopoulos et al. (2011).

Work and vocational services compared to the AB or control group, the intervention was not shown to have any effect on employment outcomes.

Subgroup analysis was conducted by type of impairment and age. For participants with mental health impairments, both treatment groups experienced statistically insignificant decreases in the share ever employed, compared to the control group. However, participants with "other impairments" increased employment, with the AB Plus group having an impact of +3.4 percentage points (the AB group had a +3.8 percentage point change, though not statistically significant).

Promoting Opportunity Demonstration (POD). Building off BOND and BOPD, POD tests new benefit offset rules that simplify work incentives to promote employment and reduce administrative complexity. The simplified POD rules eliminated the TWP and Grace Period, and they used a uniform benefit offset formula. The demonstration varied rules about SSDI benefit termination across two treatment arms. The first treatment group (T1) could not have their benefits terminated for work, whereas the second treatment group (T2) could have their benefits terminated if participants were in full offset for 12 consecutive months. Implemented in eight states, POD enrolled 10,070 participants, randomly assigning 3,343 participants to T1 and 3,357 participants to T2. The simplified POD rules appear to have driven higher use of the benefit offset, with 24 percent of POD treatment participants (both T1 and T2) using the offset one year after enrollment, compared to only 7 percent of BOND treatment participants.

Impact estimates discussed here come from POD's interim report and compare the control group to pooled treatment groups. As of one year after enrollment, POD had no detectable impact on any of the primary outcomes. With one-quarter of treatment group members using the benefit offset, POD rules should mechanically increase benefit payments among some treatment group members. The lack of detectable impacts suggests that increases in benefit payments were offset by decreases. POD also examined various subgroup impacts, including those by age, education, employment status, and impairment type. There were no statistically significant differences within or across groups.

Demonstrations Focused on SSI Recipients

Two demonstrations that focused on SSI recipients—STETS and TETD—examined program impacts by subgroups. The subgroup-specific impact estimates for these two demonstrations are summarized in Exhibit 7.3 below.

Structured Training and Employment Transitional Services (STETS)[17]

Testing the impact of transitional employment services, the STETS demonstration offered training and transitional job placement services to young people (ages 18–24)

[17] Discussion based on Kerachsky et al. (1985).

with intellectual disability (IQ scores between 40 and 80) and limited prior work experience.[18] Funded by the US Department of Labor, the STETS demonstration implemented a transitional employment model that consisted of three phases: Phase 1 provided training and support in a low-stress work environment; Phase 2 transitioned participants to on-the-job training at local businesses in a regular work environment; Phase 3 consisted of follow-up services to those who had transitioned to competitive jobs. Program participation in Phases 1 and 2 was expected to last for roughly 12 months.

STETS operated from fall 1981 through December 1983 and implemented the transitional employment model in five locations—Cincinnati, OH; Los Angeles, CA; New York, NY; St. Paul, MN; and Tucson, AZ. The STETS demonstration enrolled 437 participants and randomly assigned 226 participants to the treatment group and the remaining 211 participants to the control group. The primary outcomes for the demonstration were employment, income, SSI receipt, and service use.

For the full sample, the evaluation of STETS found no impact of the treatment (transitional employment services) on weekly personal income of young SSI recipients with intellectual disability. It did find an increase in the fraction of recipients working in regular competitive jobs in the labor market (as opposed to working in any job, which includes training jobs that were part of the treatment). There was no detectable effect on SSI benefits received.

Subgroup analyses were conducted for the two primary outcomes—weekly income and average monthly SSDI/SSI benefits—for a range of subgroups shown in Exhibit 7.3. With respect to earnings, the subgroup analysis indicates that the treatment increased weekly income among individuals with moderate intellectual disability by $43.40, compared to the control group. It had no effect for those with borderline or mild intellectual disability. With respect to receipt of benefits, the treatment reduced receipt of SSDI/SSI among those who received other transfers (including any cash transfers and Medicaid) or no transfers at baseline by $60 monthly, compared to the control group.

Impacts on the employment outcome—percentage employed in a regular job—occurred for subgroups defined by race/ethnicity. Hispanic treatment group members experienced a 22.6 percentage point increase in their employment in a regular job, and the White non-Hispanic/Other subgroup experienced a 10.5 percentage point increase. There was not a detectable impact for Black non-Hispanic group members (their roughly 10 percentage point increase was not statistically different from zero).

[18] Though the demonstration's target population exhibited high dependence on others (measured through living arrangements), only one-third of participants were receiving either SSI or SSDI benefits.

Exhibit 7.3. Summary of Subgroup Analysis of Demonstrations Focused on SSI Recipients

Panel A: Earnings						
Age	Education	Employment Status (prior work experience)	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	Gender	Race/Ethnicity
STETS: Average Weekly Personal Income						
Age <22	Enrolled	Regular job lasting	Borderline	SSI/SSDI transfer	Men	Black non-Hispanic
T: $69	T: $67	3+ months	T: $76	T: $98	T: $80	T: $63
C: $61	C: $59	T: $85	C: $58	C: $73	C: $63	C: $56
Impact: $8	Impact: $8	C: $84	Impact: $18	Impact: $25*	Impact: $17*	Impact: $8
Age ≥22	Not enrolled	Impact: <$1	Mild	Other transfer	Women	Hispanic
T: $78	T: $74	Any job lasting	T: $66	T: $70	T: $60	T: $87
C: $66	C: $64	3+ months	C: $67	C: $60	C: $61	C: $65
Impact: $12	Impact: $10	T: $72	Impact: –$1	Impact: $10	Impact: –$1	Impact: $22
Diff: n.r.	Diff: n.r.	C: $62	Moderate	No transfer	Diff: n.r.	White non-Hispanic
		Impact: $10	T: $90	T: $46		and Other
		Other	C: $47	C: $54		T: $72
		T: $68	Impact: $43*	Impact: –$8		C: $65
		C: $56	Diff: n.r.	Diff: n.r.		Impact: $7
		Impact: $11				Diff: n.r.
		Diff: n.r.				

Age	Education	Employment Status (prior work experience)	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	Gender	Race/Ethnicity
TETD: Average Earnings after First 24 Months						
Age <22	N/A	Regular job	IQ score <40	Received SSI	Men	Black non-Hispanic
T: $2,847		T: $4,153	T: $1,939	T: $3,315	T: $3,092	T: $3,173
C: $1,635		C: $3,186	C: $710	C: $1,389	C: $1,430	C: $1,180
Impact: $1,212*		Impact: $967	Impact: $1,229	Impact: $1,926*	Impact: $1,662*	Impact: $1,993*
Age ≥22		Training/volunteer job	IQ score 40-54	Did not receive	Women	White non-Hispanic and Other
T: $3,217		T: $2,934	T: $2,961	T: $3,052	T: $3,196	T: $3,102
C: $1,533		C: $904	C: $1,655	C: $1,627	C: $1,738	C: $1,699
Impact: $1,684*		Impact: $2,030*	Impact: $1,306*	Impact: $1,425*	Impact: $1,458*	Impact: $1,403*
Diff: n.r.		Sheltered workshop	IQ score 55-70	Diff: n.r.	Diff: n.r.	Diff: n.r.
		T: $3,758	T: $3,125			
		C: $1,989	C: $1,557			
		Impact: $1,769*	Impact: $1,568*			
		Other job	IQ score >70			
		T: $2,984	T: $4,108			
		C: $1,109	C: $1,296			
		Impact: $1,874*	Impact: $2,811*			
		No job	Diff: n.r.			
		T: $2,206				
		C: $905				
		Impact: $1,301*				
		Diff: n.r.				

Panel B: SSI Benefits						
Age	Education	Employment Status (prior work experience)	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	Gender	Race/Ethnicity
STETS: Average Monthly Income from SSI or SSDI						
Age <22	Enrolled	Regular job lasting	Borderline	SSI/SSDI transfer	Men	Black
T: $86	T: $117	3+ months	T: $90	T: $203	T: $91	T: $67
C: $112	C: $132	T: $96	C: $83	C: $181	C: $134	C: $93
Impact: –$26	Impact: –$15	C: $141	Impact: $6	Impact: $23	Impact: –$42*	Impact: –$25
Age ≥22	Not enrolled	Impact: –$45	Mild	Other transfer	Women	Hispanic
T: $129	T: $91	Other job lasting	T: $98	T: $54	T: $109	T: $94
C: $138	C: $115	3+ months	C: $144	C: $114	C: $102	C: $95
Impact: –$9	Impact: –$23	T: $107	Impact: –$46*	Impact: –$60*	Impact: $7	Impact: <(–$1)
Diff: n.r.	Diff: n.r.	C: $88	Moderate	No transfer	Diff: n.r.	White and other
		Impact: $19	T: $132	T: $36		T: $118
		Other	C: $85	C: $64		C: $142
		T: $95	Impact: $47	Impact: –$28		Impact: –$24
		C: $136	Diff: n.r.	Diff: n.r.		Diff: n.r.
		Impact: –$40*				
		Diff: n.r.				

Age	Education	Employment Status (prior work experience)	Type of Health Impairment	SSDI vs. SSI vs. Concurrent	Gender	Race/Ethnicity
STETS: Average Monthly Income from SSI or SSDI						
Age <22	Enrolled	Regular job lasting	Borderline	SSI/SSDI transfer	Men	Black non-Hispanic
T: $86	T: $117	3+ months	T: $90	T: $203	T: $91	T: $67
C: $112	C: $132	T: $96	C: $83	C: $181	C: $134	C: $93
Impact: –$26	Impact: –$15	C: $141	Impact: $6	Impact: $23	Impact: –$42*	Impact: –$25
Age ≥22	Not enrolled	Impact: –$45	Mild	Other transfer	Women	Hispanic
T: $129	T: $91	Other job lasting	T: $98	T: $54	T: $109	T: $94
C: $138	C: $115	3+ months	C: $144	C: $114	C: $102	C: $95
Impact: –$9	Impact: –$23	T: $107	Impact: –$46*	Impact: –$60*	Impact: $7	Impact: <(–$1)
Diff: n.r.	Diff: n.r.	C: $88	Moderate	No transfer	Diff: n.r.	White non-Hispanic
		Impact: $19	T: $132	T: $36		and Other
		Other	C: $85	C: $64		T: $118
		T: $95	Impact: $47	Impact: –$28		C: $142
		C: $136	Diff: n.r.	Diff: n.r.		Impact: –$24
		Impact: –$40*				Diff: n.r.
		Diff: n.r.				

Notes: Due to rounding, values that were computed at greater levels of precision may appear not to sum as whole numbers.
The "Impact" is the treatment-control difference; the "Diff" is the between-group difference in subgroup impacts.
n.r. indicates that the evaluation did not report the value (but in some cases, it did report whether the value was statistically significantly different from zero and reported and is noted).
n.s. indicates that the diff value, not reported, is not statistically significantly different from zero.
Statistical significance is indicated (per the specific evaluation's conventions) as follows:
* Impact is statistically significantly different from zero
† Diff (impact differential between subgroups) is statistically significantly different from zero.

STETS also saw substantial differences across treatment sites. For example, the treatment group in St. Paul (MN) experienced a 23 percentage point increase in employment in a regular job. The treatment group in Cincinnati (OH) saw a nearly 15 percentage point increase in employment in any paid job. On earnings from regular jobs, the intervention increased earnings in all sites, but in only St. Paul and Los Angeles (CA) were those impacts statistically significant, at around $25 per week.

Transitional Employment Training Demonstration (TETD)[19]

The goal of TETD was to assess whether transitional employment services are effective at supporting SSI recipients with intellectual disability to gain economic self-sufficiency and maintain employment in competitive jobs. To test the impact of transitional employment services, TETD solicited volunteers from SSI recipients with intellectual disability ages 18–40 in 13 areas where the demonstration was taking place. Of 745 applicants, TETD randomly assigned participants into the treatment group (N=375) or the control group (N=370). The treatment group was offered access to transitional employment services, which consisted of three core services: placement in a "competitive" job; specialized on-the-job training; and post-placement support for job retention.

In a sign that the treatment was effective, the demonstration was able to place two-thirds of the treatment group in jobs, with half of them (or one-third of the treatment group) maintaining employment in the jobs, a rate consistent with other transitional employment programs. These services were time limited and were available to participants for only one year after enrollment in the study. The control group could not access these services but could access other services generally provided to SSI recipients.

The primary outcomes for the demonstration were employment, income, SSI receipt, and service use. The total earnings of the treatment group nearly doubled compared to the control group's, with the treatment group averaging slightly more than $3,100 and the control group averaging less than $1,600 during the first 24 months after enrollment. The increase in earnings is due to an increase in employment, with the share of the treatment group receiving any earnings being 18 percentage points greater than that of the control group. Though the demonstration increased earnings, it led to only small reductions in the average SSI benefits payment to treatment group members, with total SSI payments dropping by 4 percent ($266) over the 24-month period.

TETD engaged in extensive analysis of potential differences in treatment outcomes by subgroup. The subgroups were based on demographic (age, race/ethnicity, gender) and personal characteristics (IQ, motivation, physical ability); prior experiences (living arrangements, work experience, Social Security benefit receipt); and program services received. STETS had found that treatment was effective

[19] Discussion based on Thornton and Decker (1989).

at raising income for participants with more severe intellectual disability, defined as IQ scores between 36 and 51. By contrast, TETD found that participants with less severe intellectual disability had higher earnings in the treatment group than in the control group. Only participants with an IQ score below 40 did not experience increases in earnings, with the treatment raising earnings for all the other participants.

Prior work experience before TETD enrollment also produced different impacts across groups. The impact on earnings for participants with less work experience was large, whereas that was not the case for participants with prior work experience in a regular job. This finding indicates the demonstration might have effectively transitioned participants without formal work experience into the mainstream workforce and increased their earnings.

Finally, although both young and old participants in the treatment group saw increased earnings, those older than age 22 experienced a larger increase in earnings than younger participants did. TETD did not provide a test of whether the subgroups' impacts were statistically significantly different from each other.

Like STETS, TETD exhibited some differences in the estimated treatment effects by site. Some sites implementing its intervention produced better outcomes than others did. Three project sites were particularly successful at increasing average earnings of treatment group members. One of them more than doubled the average earnings of the treatment group (an increase of about $2,000 in annual earnings) over the three-year period studied. That program placed participants in light manufacturing and assembly jobs.

Potential Insights and Practical Considerations

Potential Insights from Subgroup Impact Estimates

The demonstrations reviewed here vary in the outcomes studied and subgroups considered, partly due to the interventions and populations studied, partly due to practical considerations further discussed below. Nevertheless, some overarching insights emerge.

Overall, it is apparent that in many instances, the subgroup analysis reaffirms the main impact estimates. When impacts of the intervention are present, these are often reflected across subgroups (e.g., BOND for SSDI benefits; MHTS and TETD for earnings). Similarly, when no impact of the intervention is found for the overall population studied, most subgroup estimates are not statistically significantly different from zero, as well (e.g., BOND, BOPD, POD, and STETS for earnings; MHTS for SSDI benefits).

Yet, in some important cases, subgroup estimates diverge from the main impact estimates. For example, in BOND, the main impact estimate on earnings is not significantly different from zero. However, in Stage 2 of BOND, participants with "other impairments" (other than major affective disorder or a musculoskeletal disorder) and younger beneficiaries experienced increases in earnings. Similar patterns

occur for other demonstrations for which the main impact estimate is not statistically significantly different from zero (e.g., BOPD, STETS). As further discussed in the next main section, Recent Advances, such findings can be useful for deciding whether additional research is warranted to further explore variants of the intervention for the relevant subgroups.

In contrast, in MHTS, most subgroups experienced the overall increase in earnings, with the exception of younger workers, who did not experience an increase. Similar exceptions are observed for other demonstrations for which the main impact estimate is statistically significantly different from zero (e.g., Project NetWork, TETD). These results can be helpful for diagnostic purposes, either for improving certain aspects of the intervention or for considering separate programs for specific subgroups (e.g., for younger beneficiaries).

Despite the inherent variation across studies, there appear to be some broad common patterns across demonstrations beyond practical considerations discussed in the next subsection. In particular, there are recurring differences across age groups (e.g., BOND, MHTS, TETD) and across impairment groups (e.g., BOND, MHTS, Project NetWork, STETS, TETD). SSA has recognized these differences in impacts by age and impairment groups and has conducted demonstrations that focus on specific subgroups of beneficiaries, such as MHTS (and the current Supported Employment Demonstration) for those with impairments related to mental health, or the Youth Transition Demonstration (and the current Promoting Readiness of Minors in SSI demonstration and the Ohio Direct Referral Demonstration) for younger SSI beneficiaries.[20]

There are also some insights arising from findings for specific subgroups in particular demonstrations. For example, the only demonstrations to analyze separate impacts by race/ethnicity were STETS and TETD. In the case of STETS, the results point to a lack of statistical power to isolate subgroup effects for Black non-Hispanic beneficiaries, even though the impact estimate is of similar order of magnitude as other race/ethnic groups analyzed. Other demonstrations have pointed to the potential role of receipt of other (non-SSA) programs, such as "other transfers" (STETS) or Medicaid buy-in (BOPD).

Finally, in several of the demonstrations, there appears to be variation in the treatment effects across program sites participating in the evaluations. This was true for STETS and TETD, but also for BOND and the other demonstrations focused on SSDI or concurrent beneficiaries. Though this variation typically poses some challenges in interpretation, it is not uncommon in large social demonstrations, and some guidance for data collection and analysis has emerged from the past literature on social experiments (e.g., Rothstein and von Wachter 2017).

[20] See Chapters 5 and 6 for more information on those.

Practical Considerations from Subgroup Impact Estimates

In terms of practical considerations arising from the discussion of subgroup impact estimates in the earlier subsection "Demonstrations Focused on SSDI and Concurrent Beneficiaries," it might be worthwhile to consider a broad set of common standards for the definition of subgroups, the choice of outcomes, and statistical specifications.

Clearly, the observed differences in outcomes and subgroups between demonstrations shown in Exhibits 7.2 and 7.3 arise partly from differences in the interventions and populations studied. For example, a demonstration focused on, say, beneficiaries who are younger (such as STETS) or beneficiaries with mental health impairment (such as MHTS) might benefit from subgroup definitions different from those for a demonstration focused on a broader population (such as BOND).

Yet, some of the variation across studies appears relatively minor. For that, it might be worth settling on a template of default choices for certain common subgroups that studies can use as a benchmark and modify as needed. To obtain a broader and more consistent coverage of subgroups, such a template could also be used to signal which subgroup categories should be considered for inclusion in future demonstrations (e.g., such as race/ethnicity and gender).

Similarly, the variation in outcomes used appears to be partly due to choices made in the study (e.g., focus on short- versus longer-term earnings outcomes), and partly due to data limitations. For example, BOND and Project NetWork, among others, both use SSA's Master Earnings File to estimate total earnings 2011–2015 and average annual earnings, respectively. The BOND report notes that these data are available only by calendar year and are not precisely aligned with randomization (conducted in May 2011), potentially inflating estimated earnings. In contrast, MHTS estimates earnings using survey data. In future demonstrations, it might be valuable to institute a common set of earnings and benefit metrics based on SSA administrative records available in the same fashion to all demonstrations. To complement information from administrative records while maintaining comparability between studies and between different earnings measures, it is also worth considering a standardized template for survey-based earnings and income measures.

The majority of studies reviewed in this chapter did not systematically report information that would allow us to assess whether subgroup-specific impacts are statistically different from one another within a category (e.g., men's versus women's impacts, within gender). It is worth considering requiring reporting of such tests for all subgroup impacts for future demonstrations.[21] Another aspect is that subgroup estimates come from separate, subgroup-specific treatment-control comparisons. If subgroup characteristics are correlated in the population, we risk attributing subgroup

[21] More-recent demonstrations seem more likely to include relevant information, so there may be some acknowledgment of the need for this already. However, it does not appear to be consistent and it could be useful to standardize the reporting of these estimates.

effects to one subgroup because of correlation with another subgroup. For example, suppose that both older beneficiaries and less-educated beneficiaries are found to be less responsive to a benefit offset. Though it might be that both characteristics matter independently in determining the effect of the intervention, it could also be that the effect for older beneficiaries arises if most older beneficiaries are less educated. Similarly, it could be that a subgroup appears to matter only because it is correlated with another subgroup. For example, older individuals tend to have more labor force experience than do younger individuals. Hence, it might be that only experience matters for the outcome of an intervention, but that the age effect when examined alone is found to be statistically significant because of the correlation of age with experience.

A single regression model in which the treatment indicator is interacted with each relevant subgroup category would be able to show the impact of potential correlations among subgroups. The interpretation of the coefficients and their standard errors depends on the specification of the model. For example, *if a main effect for the treatment is included*, one has to exclude one subgroup-treatment interaction from each category.[22] In that case, the main effect for the treatment measures the impact for the excluded category. The coefficients on the subgroup indicators measure the *difference* in the effect of the treatment relative to the excluded group (whose effect is measured by the main treatment indicator), netting out any effects arising from correlation with other subgroups.

For example, if we have three categories—say, gender, a binary age-group indicator (e.g., older versus younger), and a binary indicator for recent work experience (e.g., worked in past five years versus did not work in past five years)—and we excluded the subgroup-treatment indicators for men, younger workers, and no recent work experience, then the main treatment indicator would capture the effect for younger male beneficiaries with no recent employment (the excluded group). The coefficient on the interaction between the treatment effect and an indicator for older beneficiaries would show how the effect for older beneficiaries differs on average from the excluded group, *holding constant* differences in the effect of treatment that could arise because gender and work experience are correlated with age (e.g., if older beneficiaries were more likely to be men and have more work experience). If the subgroup effect for age, considered alone, mattered only because of the correlation of age with recent employment experience, then the coefficient in the interacted model should not be statistically significantly different from zero. In other words, conveniently the standard errors on the interaction effects in the model with multiple interactions can be used to construct test statistics for assessing whether a particular subgroup-specific impact estimate is different from the main effect, *conditional* on inclusion of the remaining subgroup interactions.

---

[22] This assumes, as is commonly the case, that subgroups within a category completely describe the population (e.g., education less than high school, equal to a high school diploma, or more than a high school diploma), such that the indicators for the subgroups add up to the constant term.

If *we do not include the main effect*, then we can include subgroup-treatment interactions for all subgroups to be analyzed. The coefficient on each subgroup-treatment indicator then measures the effect of the treatment on the particular subgroup, holding constant differences in the effect of treatment arising due to correlation with other subgroups. Continuing our stylized example above, if the treatment effect on older individuals is found to be statistically significantly different from zero in this model, then older workers experience a different treatment impact than younger workers, even holding constant the level of work experience.

RECENT ADVANCES IN ESTIMATING HETEROGENEOUS TREATMENT EFFECTS

Clearly, for large populations such as SSDI or SSI participants, the likelihood is high that there is heterogeneity in the response to the particular intervention. Though under random sampling the main treatment effects yield the average treatment effect (ATE) in the relevant population, the intervention might be working better for some groups within that population than for others. For example, this was the case for younger individuals in MHTS or for less-educated workers in Stage 2 of BOND. This section first summarizes existing and new statistical approaches to uncover treatment effect heterogeneity. Then it discusses under what circumstances estimates of treatment effect heterogeneity could be used in future evaluations.

Statistical Approaches to Estimate Extent of Treatment Effect Heterogeneity

The traditional approach of assessing treatment effect heterogeneity in evaluation research is to pursue a limited number of group-level contrasts that are pre-specified in the evaluation's analysis plan. Pre-specification solves the potential bias that could arise if researchers wait to choose contrasts based on the observed outcomes of the evaluation. Pursuing a limited number of contrasts also avoids the risk of finding statistically significant contrasts purely by chance. The traditional approach can yield important insights into treatment effect heterogeneity among key groups relevant to the particular program (see also the discussion in the earlier subsection "Other SSDI-Related Demonstrations"). Yet, by limiting the analysis of heterogeneity to a handful of covariates, the traditional approach might not be able to effectively isolate the relevant margins among which heterogeneity could occur. Indeed, the preceding section, "Discussion of Heterogeneity in Estimates for Demonstration Outcomes," has shown that analyses of even a limited number of subgroups can result in a large number

of potentially imprecisely estimated contrasts even in demonstrations with comparatively very large sample sizes such as BOND.[23]

A growing literature in statistics and economics has devised a range of approaches tailored to the scenario in which the effect of treatments is heterogeneous across individuals. Here I will briefly discuss two broad categories of approaches, one that involves modifying the experiment ex-ante, and another one that affects the way the experimental data are used ex-post.

These two approaches build on important developments in statistics that clarified the assumptions needed and the interpretation of impact estimates in an environment with heterogeneous treatment effects and non-compliance (i.e., when not all individuals take up an offered treatment). Similar considerations apply in an environment where individuals are asked to volunteer for a program, as is the case for the SSA demonstrations. Though a detailed summary of these developments goes beyond the scope of this chapter, the key insight was that individuals could self-select into the program based on their perceived valuation of the treatment, with this self-selection often taken to be the relevant treatment effect (e.g., Angrist, Imbens, and Rubin 1996; Frangakis and Rubin 2002). This insight implies that researchers must make assumptions about individuals' choices after they are exposed to the intervention studied, and it motivates the role of the probability of take-up (the propensity score) discussed below. The literature has shown that a judicious use of such assumptions can yield insights into the nature of heterogeneous treatment effects, and in some cases into the nature of the treatment itself.[24]

These approaches are an important part of the experimental researcher's tool kit and can be helpful in understanding core dimensions of heterogeneity in treatment effects. Their application does vary on a case-by-case basis, they are not meant to deliver fine-grained estimates of individual or group-specific treatment effects, and these approaches are now well-covered elsewhere (e.g., Imbens and Rubin 2015). Ultimately, it is an important empirical question whether the recent methods based on innovations in data science and machine learnings discussed here yield substantive empirical improvements to the standard approach to subgroup analysis or those arising from a deeper understanding program choice.

---

[23] Another potential drawback of standard subgroup analysis is that due to the smaller sample sizes in the subgroups, such analyses are more likely to be capable only of finding evidence of statistically significant effects that are larger than the main effect, rather than vice versa. (Given that significance tests rely on a comparison of an estimate relative to its standard error, and smaller sample sizes imply on average larger standard errors, only larger effect sizes will be found to be statistically significantly different from zero in subgroup analysis.) Yet, for diagnostic purposes, cases in which treatment effects are smaller in absolute value than the main impact estimates are important, as well.

[24] Page et al. (2015) summarize the "principal stratification" approach, as it is called in the statistics literature; and Peck (2013) summarizes the "endogenous subgroup analysis" approach, as it is called in the program evaluation literature.

Ex-Ante Approaches to Improve Estimates of Treatment Effect Heterogeneity

Rothstein and von Wachter (2017) discuss a variety of ways in which experiments could be structured to uncover differences in treatment effects along particular dimensions. Suppose, for example, a continuous measure was available of beneficiaries' (potentially unobservable) underlying earnings potential. For example, the average of an individual's prior earnings is often a strong predictor of future earnings.[25] Then randomization could occur within stratified groups based on predicted earnings potential. Such an approach would, as before, allow estimation of the ATE without a loss in power. In addition, if the sample sizes in each stratum on that measure are appropriately chosen, it would allow a more targeted and interpretable analysis of heterogeneity in treatment effects by differences in labor supply potential.[26] For example, the National Income Tax experiments were stratified based on prior earnings, which can be seen as a predictor for the responsiveness to different income tax rates (Ashenfelter and Plant 1990). An additional advantage of this approach is that it is likely to raise the probability of detecting treatment effects of inducements to return to employment if individuals with the highest estimated earnings potential are also most responsive to such inducements.[27]

Another promising candidate for such cross-classified experiments is a measure of the probability of taking up the offered treatment. In experiments when not all members of the treatment group take up the treatment, research has shown that under some conditions the probability to take up treatment (sometimes called the propensity score) can be used as an index of an individual's benefit from the treatment. Intuitively, if those individuals who benefit more from the treatment are more likely to take up the

[25] Alternatively, earnings could be predicted based on information contained in SSA's records on the beneficiary, such as the Residual Functional Capacity questionnaire, or on information from the continuing disability reviews.

[26] Without changing the overall sample size, stratifying the random assignment ensures that sample sizes of strata do not differ due to random variation.

[27] If the effect of treatment does not vary with estimated earnings potential, there is no gain from the stratification in learning about the distribution of underlying treatment effects. However, the resulting subgroups could be of interest for other reasons. For example, the effect of the treatment for low-earning individuals might be of interest in its own right. Moreover, there is also no loss, in the sense that we can still obtain the estimated ATE for all treated individuals.

treatment, then an estimate of the propensity score can be used to isolate those individuals more likely to benefit from the treatment.[28]

Hence, we could design an experiment that cross-classified the treatment using strata defined based on an estimate of the propensity score based on data for the experiment. Again, if sample sizes are kept unchanged, cross-classified randomization ensures balance in sample sizes. An increase in sample sizes for each or at least some strata might be warranted based on standard power calculations. The resulting treatment effects for each stratum would characterize differences in the underlying treatment effect in the population. If stratification by the propensity of take-up is further pursued *within* groups defined by observable characteristics, then in principle the entire distribution of treatment effects for the population can be estimated (e.g., Heckman and Vytlacil 2005).

A closely related approach would be to directly manipulate the probability of take-up of the program with an additional, cross-classified *treatment*. For example, suppose preliminary research showed that distance to the job training or transitional job services site was an impediment to treatment, and that transportation subsidies can increase participation. Then we could cross-classify the randomization of the original treatment (e.g., job training) with randomly assigned transportation subsidies of different amounts. Because the staggered transportation subsidies directly manipulate the probability of take-up, the resulting data can be used not only to better estimate heterogeneity in treatment effect, but also to better understand the effectiveness of approaches for reducing barriers to program take-up.[29]

---

[28] A key assumption is that individuals' choices to take up treatment can be represented by a single summary measure (such as their net monetary gain based on expected increase in earnings minus travel and child care costs). If this is not the case—for example, if individuals with high potential treatment effects are also those who do not understand the benefits from the program—or if the gross gain is of interest to the policymaker (e.g., the effect of treatment on earnings independent of child care or travel costs), then the use of the propensity score as an index of underlying treatment effects has to be reconsidered and potentially modified.

[29] Rothstein and von Wachter (2017) discuss another version of a cross-classified experiment that solves another common problem in the analysis of workforce training. Though often we would like to understand the program impact on an outcome that is observed *conditional* on working (e.g., hourly wages, or annual earnings conditional on working, as in MHTS), the impact estimates on such conditional outcomes are not identified in the basic experimental design because they rely on an endogenous choice—employment—that introduces a potential bias in conditional estimates. Consider then a cross-classified experiment that in addition to the treatment (e.g., a training program) manipulates the decision to work after completion of treatment (e.g., through randomized provision of commuting subsidies or some other relevant means to improve access to job sites). By manipulating the employment decision, this approach allows estimating the effect of training on hourly wages, which is often used as a measure of productivity (or on earnings conditional on working). As a result, the combined experiment allows estimating the effect of the original treatment both on employment and on wages. This is more informative than analyzing annual or even weekly earnings, because earnings reflect both labor supply and wages.

In practice, researchers have to choose how to measure the variable used for cross-classified randomization, such as earnings potential or probability of program take-up. Obtaining such a measure is a key step of the research design. Some of the approaches discussed in the next subsection, "Ex-Post Approaches to Improve Estimates of Treatment Effect Heterogeneity," can be applied to correlational data or to the quasi-experimental studies described earlier in the section "Background on Variation in Employment Potential" to obtain improved measures of earnings potential. Thereby, results from existing demonstrations could be used to guide high-level modeling choices. For example, age, prior employment, or impairment types have shown to be important predictors of employment potential in both quasi-experimental studies and demonstrations; in principle, they could be used to form a coarse ex-ante classification. Alternatively, they could be used to define higher-level groups within which a statistical algorithm provides refinements. Thereby, detailed information available at SSA, such as long earnings histories, medical determination, residual functional capacity, and place of residence, could be combined with information on indicators of local labor markets, such as the incidence of vacancies in jobs similar to the occupation previously held by a beneficiary.

From a theoretical point of view, the most accurate possible measure would be preferable, but as further discussed later (in "Practical Considerations for Use of Heterogeneous Treatment Effect Estimates"), practical considerations could lead researchers to choose variables that can be potentially observed as the program is administered by case managers. For example, the earnings potential of SSDI beneficiaries could be based on the amount of prior earnings, on their predicted earnings (e.g., based on their demographics, education, occupation, and employment experience), or on a more sophisticated measure based on an assessment of the market value of their residual functional capacity or their occupation-specific skills.[30] Similarly, the estimated probability of treatment take-up could be based on the full set of information available to SSA, on a subset that is deemed sufficiently predictive of take-up, or on additional variables that currently might not be routinely collected but were found to be predictive of program take-up in preliminary studies for the particular experiment or in analysis of data from existing demonstrations.

[30] As an alternative, we could use estimates of heterogeneity in labor supply effects to SSDI benefits from studies that seek to use random variation occurring naturally in the data as discussed earlier in "Discussion of Heterogeneity in Estimates for Demonstration Outcomes" (e.g., French and Song 2014; Hemmeter and Bailey 2016; Maestas, Mullen, and Strand 2013). Though preferable insofar as these are already estimated treatment effects from being exposed to the SSDI program in various ways (depending on the study), they might not be as suitable for diagnostic or targeting purposes because these treatment effects cannot be easily measured in the population (see the subsection "Practical Considerations for Use of Heterogeneous Treatment Effect Estimates").

Ex-Post Approaches to Improve Estimates of Treatment Effect Heterogeneity

The potential statistical issues of traditional approaches to subgroup analysis have been a motivation for algorithmic (or data-driven) approaches to exploring treatment effect heterogeneity that can be applied to existing experimental data. Several approaches have been developed that implement statistical algorithms by which the computer, not the researcher, runs through a large number of contrasts based on flexibly defined categorical groups. From that, it obtains estimates of treatment effects for these contrasts and their standard errors.

The following gives a broad overview of two groups of recent approaches.[31] Based solely on the way the treatment effect is calculated, I will refer to these as "semi-parametric" and "non-parametric" approaches. All of these approaches are designed to explore treatment effect heterogeneity among a large number of categories (e.g., age by gender by education by income class). They differ in their data requirements and in their statistical properties. Depending on the particular application, they can differ in the interpretability of the resulting subgroups, as well.

Semi-parametric. What I call "semi-parametric" approaches can be viewed as an extension of the traditional approach. As discussed regarding the traditional approach in the opening of this "Statistical Approaches" section, a semi-parametric approach can be formulated in a standard linear regression framework in which the desired outcome is the outcome variable, and the treatment dummy interacted with the pre-specified subgroup indicators are the main control variables.[32] The coefficients on the interactions represent the group-specific treatment effects (or differences in the effects with respect to an omitted category, if a constant term is included).[33] Recent publications have extended this approach to allow estimation of treatment effects that are a fully flexible function of all the covariates, rather than just a function of group indicators. Various statistical approaches are used to automatically choose the most relevant treatment effect differences among a potentially large number of contrasts. Because no restriction is placed on how the treatment effects depend on the observed covariates, this approach can be viewed as non-parametric estimation of how treatment effects vary with covariates.

A version of this semi-parametric approach shows intuitively how estimates of individual-level treatment effects can be modeled flexibly as a function of covariates.

---

[31] See Peck (2005) for an example of an early approach to data-driven choices of subgroup contrasts that uses statistical methods to obtain a limited number of clusters of similar observations in the population that can be used instead of or in addition to traditional groups for obtaining subgroup contrasts.

[32] In the case of an experimental evaluation, additional control variables are sometimes added for precision, but I ignore these here for simplicity.

[33] Without a constant term, the coefficients simply measure the treatment effects by group. Alternatively, if the coefficients on interactions are constrained to sum to zero, they represent the difference with respect to the mean treatment effect.

Consider drawing for each treated subject a statistical twin from the pool of control participants. This can be done, for example, by choosing two individuals who have the same estimated probability of treatment (the propensity score discussed earlier in the "Ex-ante Approaches" subsection). The difference in outcomes between these two individuals can be viewed as a coarse estimate of the individual treatment effect. This individual treatment effect can then be regressed on a flexible functional form of the covariates, as in any standard semi-parametric or non-parametric regression. Standard machine learning and related approaches (e.g., least absolute shrinkage and selection operator, or LASSO) can be used to choose among a potentially large number of covariates.[34]

The resulting estimates can be used to predict a treatment effect for each individual based on their covariates' values (something also referred to as conditional ATE, or CATE, because it is conditional on the given individual's observed covariates). The distribution of estimated CATEs can be analyzed as a whole or for different subgroups to assess the estimated degree of heterogeneity among treatment effects. The CATEs can also be used, in principle, to re-calculate the overall treatment effect, as if different populations of individuals had received treatment (e.g., under different program rules or different outreach strategies).[35] Such an exercise can be a useful diagnostic device to assess how much potential room for improvement there might be for better targeting the intervention.

Non-parametric. The second, non-parametric approach dispenses with the linear regression framework and uses the patterns in the data directly to isolate those groups that have the largest differences in treatment effects. This is done by recursively partitioning the data into those groups that exhibit the largest difference in treatment effects. The algorithms used for the stepwise partitioning (called "causal trees") can more flexibly search over different combinations in the data than the semi-parametric approach can, and hence might be more likely to isolate salient differences in treatment effects. The final product is again an estimate of a CATE for each individual based on their covariates. These estimates can be used for analysis of treatment effect heterogeneity in the same way as the results from semi-parametric estimation can.[36]

---

[34] The same result can be achieved by separately modeling the counterfactual outcomes under treatment and non-treatment as a function of covariates, and then constructing individual treatment effects as a difference of each individual's estimated counterfactual outcomes (e.g., Foster, Taylor, and Ruberg 2011).

[35] This is pursued, for example, by Knaus, Lechner, and Strittmatter (2020), who use the semi-parametric approach to estimate the effect of job search programs in Switzerland.

[36] The CATEs obtained from this method have been shown to have desirable statistical properties (Athey and Imbens 2016).

Choosing a Statistical Approach

Different approaches differ in their data requirements, ease of implementation, and statistical properties. For example, the semi-parametric approaches can be used to choose among a large number of covariates with respect to the total sample size. The non-parametric approaches discussed can also deal with a large number of covariates but tend to require a larger number of observations. In both cases, a higher number of observations and a higher coverage of observations along possible dimensions of heterogeneity will lead to more accurate estimates of the underlying heterogeneity in treatment effects.

Though the algorithms replace the researchers' potentially confounding choices of dimensions of heterogeneity, it is important to bear in mind that researchers will still have to specify aspects of the analysis that can influence the final outcome (e.g., the smoothing parameters of the machine learning algorithms, or the partitioning of the data into training and estimation samples). As this again risks introducing researcher-induced variation in outcomes, sensitivity analyses along the relevant margins are an important step in the implementation of these measures.

Another feature of all of the statistical approaches mentioned is that the result might or might not be readily interpretable. Interpretability might not be required if, for example, the main goal is to predict which individuals will most benefit from treatment to better assign treatment directly based on the estimated CATEs. However, as further discussed next, in some cases, interpretability is a desired feature of the analysis, for example, if the analyses are meant to inform the understanding of the treatment more broadly or if the results are meant to be used to generate new assignment mechanisms. However, the researcher could modify the type of covariates used for the heterogeneity analysis and see whether a reduced set of covariates is still able to provide a good fit to the observed heterogeneity in treatment outcomes.

Practical Considerations for Use of Heterogeneous Treatment Effect Estimates

Several potential practical and ethical aspects arise when considering the potential use of estimated heterogeneous treatment effects in practice. In the following, I will briefly discuss such aspects for three potential use cases for estimated heterogeneous treatment effects: diagnostic purposes, intervention targeting, and intervention evaluation.

For Diagnostic Purposes

An inspection of the estimated CATE discussed above in the "Statistical Approaches" section can indicate for which individuals the intervention might not be working as well as for others. Systematic pattern in the CATEs might further give clues as to the underlying sources of the differences. One important question that arises is what constitutes the underlying sources for heterogeneous outcomes. In the context

of the demonstrations discussed in this chapter, treatment effect heterogeneity is likely to arise from individual characteristics that affect the ability or desire to work. These characteristics could lead to differences in the effect of the treatment, in the sense that as designed, the treatment does not "work," or works less for some individuals than for others.

Yet, treatment effect heterogeneity could also arise from other sources than the nature of individuals and of the treatment. It has often been observed that different program sites exhibit different average treatment outcomes. Such "site effects" can arise due to differences in how the program is implemented or administered, which in turn can influence the composition of individuals being served.[37] Heterogeneity in outcome could also arise from differences in participants' understanding of the intervention, interactions with caseworkers or teachers, or access to other employment supports, just to name a few.

The analysis of the estimated CATEs can be useful in this process, as different variables used for estimation can reflect some of the potential underlying sources of heterogeneity. Follow-up research (e.g., quantitative or qualitative surveys, focus groups, or reviews of program fidelity and program process, among others) is likely needed to complement the quantitative analysis. Besides being directly useful to the problem at hand, this information can also help to provide information on what data can be collected in future demonstrations to obtain more informative estimates of CATE.

For Intervention Targeting

Heterogeneous treatment effects can also in principle be used for better targeting the intervention to those individuals who will benefit the most. Whether this is desirable ultimately could be a decision based on a range of factors outside of the realm of quantitative analysis. However, in principle, the estimated CATEs, possibly together with estimates of the cost of treating individual participants, could be used to generate lists of individuals for whom the estimated cost-benefit of the intervention is particularly high. Such a list could be prioritized for treatment or for proactive outreach to take up treatment. As further discussed below, when pursuing such a strategy, it is important to consider potential risk of propagating pre-existing biases in program access or success.

From a practical point of view, this targeting requires that the information used to calculate the CATE be available to administer the program on daily basis. For example, if mostly based on administrative data that are accessible in real time, the CATE could be calculated based on up-to-date information for use by a caseworker or an outreach team. In other instances, the CATE might be too complex; that is, based

[37] The composition of individuals can differ across sites for other reasons (e.g., differences in local populations), and controlling for such differences when evaluating site effects can be a useful step in diagnosing potential sources of site differences.

on information not easily accessible to SSA (e.g., if it was based on survey data collected for a demonstration) or based on administrative data not readily available to caseworkers on the ground. In this not uncommon scenario, the research team can assess whether it is feasible to generate informative estimates of the CATE based on a subset of variables that are more readily available to relevant caseworkers or that could be collected at low cost as part of a modified intake process (e.g., in the form of a short intake or targeting questionnaire).

A potential concern for targeting is that the estimated CATEs are subject to sampling error. If a fixed cutoff for a CATE were to be used for prioritizing individuals for treatment, this would lead both to false negatives (lower prioritization of individuals whose true CATE is above the cutoff) and to false positives (higher prioritization of individuals whose true CATE is below the cutoff). Depending on estimates of the size of the sampling error and a given cost of making an error, the cutoff could be sufficiently relaxed to avoid making mistakes that are deemed too costly. More generally, given variability from sampling, together with normal uncertainty regarding the correct statistical model or the CATEs, appropriate caution is advised when using estimated CATEs to exclude individuals from treatment outright.[38]

Another concern is that estimated CATEs inadvertently propagate pre-existing biases or discrimination. For example, if in the past no effect was found on a subgroup because of the influence of racial or ethnic discrimination, then using the resulting CATEs into future targeting would risk propagating that same discrimination. This is a well-known problem in the literature using machine learning algorithms to predict future outcomes. Such "predictive analytics" can be useful in its own right for targeting by helping to identify which individuals are more likely to take up an intervention. Predictive analytics also can be used to generate probabilities that are useful for stratified analysis (e.g., predicting the probability of a particular beneficiary working above the SGA level). Approaches used in the literature on predictive analytics for detecting bias, together with institutional and qualitative information on the intervention studied, can be used to prevent propagating any biases in treatment assignment that might be introduced by the use of estimated CATEs.

For Intervention Evaluation

Finally, the estimated CATEs could be used to inform the overall evaluation of the viability of a tested intervention. Though not the sole decisive factor, commonly estimated ATEs of an intervention play an important role in its evaluation. The

[38] There could be cases in which such an approach is reasonable. Consider the case when the expected effect of an intervention is positive. The treatment might not be considered viable for individuals with precisely estimated but large and negative CATEs, or precisely estimated but low and positive CATEs *and* high program costs.

presence of treatment effect heterogeneity makes matters more complicated, but the CATEs can provide a valuable source of information.

Consider the case of a single outcome of interest—say, the net impact of the program on individuals' disposable income (i.e., earnings plus SSDI/SSI benefits). We can represent society's valuation of the intervention as a weighted sum of treatment effects among all participants (this is sometimes called a "welfare function"), where the (welfare) weights represent the value to society of providing additional income to each individual. In this basic, yet realistic example, if society's values are equal for all individuals, heterogeneous treatment effects do not provide additional information about the program beyond the ATE. Yet, this would be a very unusual welfare function. In the United States, and in many other countries, programs are structured such that funds are transferred to lower-income individuals, reflecting that welfare weights for many social insurance programs usually decrease with income.

Suppose then that the distribution of estimated treatment effects is centered around zero impact, such that the ATE is zero as well, but that treatment effects are positive for poorer or otherwise needier individuals. In the realistic scenario of welfare weights that decrease with income, all else equal, the evaluation of such a program would be very different under ATE and CATE. The question can become more complex with multiple primary outcomes of interest. For example, in all demonstrations discussed here, measures both of earnings and of benefit receipt were primary outcomes. In this case, the estimated CATE on net income would likely receive a different weight (the welfare weight of program participants) than would the estimate of total program costs saved, which affects workers paying Social Security taxes (and hence would be evaluated at the welfare weights of, say, an average worker). Though more complex, this is not an unusual set of welfare tradeoffs, and hence estimated CATEs could be a valuable input into a broader process of evaluating the viability of proposed programs.

CONCLUSIONS

This chapter has discussed the heterogeneity of the impacts evaluated in eight demonstrations that tested the impact of work incentives and work supports for SSDI beneficiaries and SSI recipients. An analysis of heterogeneous impacts allows a better understanding of whether some participants benefit more from a particular intervention than other participants do. This, in turn, allows SSA to either target potentially expensive interventions to individuals who will benefit most or improve interventions for those participant subgroups that appear to benefit less from it.

The chapter started out by motivating the potential role of heterogeneity in beneficiaries' and recipients' employment potential, based on descriptive and non-experimental evaluations. The chapter then provided an overview of findings on heterogeneity, including a summary of the comparability of outcomes and subgroup definitions among demonstrations. It then briefly discussed potential lessons and practical implications from the discussion of heterogeneity. It then concluded with a

summary of alternative statistical approaches to address heterogeneity in the response to treatments brought forward in the recent literature, distinguishing between approaches modifying the experimental design versus those based on existing experimental data.

The chapter comes to four broad conclusions.

1. **The available results of subgroups paint a helpful but complex picture of heterogeneity of impact estimates.**

For MHTS, one of the two demonstrations with a detailed analysis of impact estimates by subgroup, the intervention (improved behavioral health services and case management) led to widespread increases in earnings for all subgroups studied, with the exception of younger beneficiaries (younger than age 35). These findings (and findings from the other demonstrations discussed in this chapter) are based on study participants drawn from a pool of volunteers and hence might not be representative for all beneficiaries with these health conditions. Yet, as long as those volunteering to participate in the demonstration are similar to those who would take up the treatment, were it offered on a larger scale, the estimated impacts are informative if the program were adopted more widely.

For BOND, the other demonstration with a systematic analysis of subgroup impacts, the subgroup analysis identifies several groups of Stage 2 volunteers for whom the treatment (benefit offset) increased earnings. This is true for younger and less-educated beneficiaries, those with some prior unemployment (at a marginal level of significance), and those with "other impairments" (other than major affective disorder or a musculoskeletal disorder, the two impairment types studied explicitly in BOND). This is notable because the overall effect on earnings was found not to be statistically different from zero. In contrast to MHTS, the BOND report shows tests for the difference within subgroups. Only the contrast between individuals with prior employment and prior unemployment almost satisfies the margin of being statistically significantly different from zero at a 10 percent level.

2. **It would be useful to improve comparability of estimated program impacts between demonstrations by adopting a core set of common definitions of subgroups and outcomes.**

Currently, SSA's demonstrations used a range of definitions of earnings and employment, making comparison between studies difficult. Similarly, demonstrations used a broad range of different subgroups, with only age, enrollment type (e.g., SSDI versus concurrent SSDI/SSI), and to some extent impairment type being comparable across studies. Settling on comparable earnings and employment measures, harmonizing subgroup definitions, and including gender, race/ethnicity, and prior education in standard subgroup analysis would be worth considering. In terms of implementation, a common set of reporting practices would be helpful (e.g., reporting of test statistics for differences of subgroup estimates).

3. **It would be useful to harmonize statistical analysis of demonstrations by reporting results from statistical tests of the difference between subgroup impacts and estimate statistical models that account for cross-group correlations.**

The review of empirical findings in this chapter suggests the lessons learned from SSA's demonstrations could be further improved by adopting a common set of standards for the statistical analysis of impact estimates by subgroups and the reporting of its results. All analyses should report statistical tests for the differences in subgroup impacts within a category (e.g., men vs. women). In addition to comparisons of differences in treatment and control group means for each subgroup, subgroup impacts should also be estimated from statistical models that include interactions of the treatment indicator with all relevant subgroups to account for potential correlations among groups in the population.

4. **It is worth exploring new ex-ante and ex-post statistical approaches to analyzing impact heterogeneity.**

Such approaches could allow analysts to assess impact heterogeneity for existing demonstrations using newly developed data science techniques. For example, in future demonstrations, stratifying randomization by comprehensive measures of work potential and oversampling some strata could improve the ability to detect impacts. Similarly, data science methods have been used to estimate treatment effects after randomization that flexibly use patterns in the data to determine relevant subgroup differences. Such approaches could lead to a richer understanding of how proposed programs affect the large and heterogeneous population of SSDI beneficiaries and SSI recipients and could allow improved service outcomes through improved targeting or intervention differentiation.

Consider the example of a demonstration of a potential new work support intervention, or an existing program that has not yet been experimentally evaluated, such as Ticket to Work. It is possible to develop approaches to target Ticket to Work, based on econometric estimates of employment potential or of the probability of taking up the program, using machine learning approaches. As discussed in this chapter, it could be possible to further improve such targeting and resulting intervention outcomes by experimentally estimating heterogeneous treatment effects. To maximize the ability to identify relevant heterogeneity in treatment effects, such a demonstration would settle, in advance, on a stratification within which randomization takes place (such as predicted employment potential based on rich administrative data available at SSA) and adjust sample sizes accordingly. An important preliminary step is research in using SSA's substantial data (containing among them information on earnings histories, medical determination, the occupation, labor market experience, and residual functional capacity) together with growing information on employment and vacancies in local labor markets to improve predictions of beneficiaries' work potential.

In addition, once the experiment has taken place, increasingly standard data science tools can be used to assess further dimensions of treatment heterogeneity and further refine these strata. The resulting distribution of estimated treatment effects (the CATEs) can then be analyzed for diagnostic purposes. Researchers can then use these treatment effects to devise approaches that can be implemented in the field to better target the intervention to those beneficiaries most likely to take up and to benefit from it. If desired, the design of the experiment and data collection can be adjusted in advance to meet the envisioned use case of the information on heterogeneity.

Finally, the review of the mixed success of the eight demonstrations to achieve broad and sustained earnings and reductions of SSDI receipt among SSDI beneficiaries (with evidence of increases in SSDI receipt in BOND) suggests that SSA should be intervening when disabled workers are younger or as soon as a new disability occurs. SSA has already begun assessing the effectiveness of early interventions (see Chapter 5 in this volume). It may be worth further exploring potential synergies between workforce interventions for hard-to-reemploy workers and those that serve partially disabled workers. Aiming to understand how partially disabled workers fare in these workforce programs, and assessing the potential costs and benefits of reintegrating the workers prior to their receipt of or application to SSDI seem worth considering in the future.

NOTE

A version of this chapter with additional detail about the demonstrations and their findings in appendix tables is available online (http://www.econ.ucla.edu/tvwachter/).

ACKNOWLEDGMENTS

I thank Alex Coblin for stellar research assistance, Landon Gibson for invaluable discussions, and Debra Goetz Engler, Kai Filion, Daniel Gubits, Jeffrey Hemmeter, Laura Peck and the chapter's discussant's Howard Goldman and Nick Hart for helpful comments on initial versions of the chapter.

Chapter 7

Comment

Howard H. Goldman
University of Maryland

My comments (regarding "An Overview of Current Results and New Methods for Estimating Heterogeneous Program Impacts," by Till von Wachter) derive from my perspective, gained over the past 40 years, as a mental health policy researcher and occasional advisor to the Social Security Administration's programs on disability. I address myself to three areas:

- Salient points from this excellent chapter with which I strongly agree.
- A point of disagreement with respect to characterization of mental impairments as less severe and people experiencing them as having greater employment potential than other groups.
- General policy lessons about disability due to mental impairments based on two demonstrations, the completed Mental Health Treatment Study (MHTS) and the Supported Employment Demonstration (SED), which is still in the field.

POINTS OF AGREEMENT

In my view, the most important general lesson on heterogeneity stated in the chapter is this: "Insights on variation in the effects of treatment for certain groups can help in better implementing interventions by informing which SSDI beneficiaries and SSI recipients might be particularly responsive to new features of a program." Recognizing heterogeneity in outcomes is important to focusing an intervention on a particular target group likely to benefit. Such targeting is important both for the effectiveness of the intervention and for recruitment of participants in a demonstration. These observations will be critical, as well, if we ever implement a program based on a demonstration. It is true that no one-size-fits-all intervention is likely to emerge, but it is still possible to expand the scope of an intervention to focus on a broader group of potential participants. For example, supported employment that follows the Individual Placement and Support (IPS) model is central to both the MHTS and the SED. Both demonstrations focus on individuals with mental impairments, but the MHTS successfully targeted individuals who receive disability benefits, whereas the SED is focusing on individuals initially denied disability benefits based on their mental impairments and comorbidities. The SED tests the dismantling of the intervention from the MHTS, comparing a Full-Service arm, which includes a nurse care manager, versus a Basic-Service arm without the nurse. It remains to be seen whether the expansion of the interventions to individuals denied benefits on initial application (with either intervention arm) is warranted. Meanwhile, in other studies in the field,

IPS is also being tested on individuals with post-traumatic stress and substance-use disorders. Conceptually, IPS could be tried with individuals with any category of impairment or mix of impairments. These demonstrations illustrate the potential benefits of focusing on the heterogeneity of outcomes.

Another important point made by the author is the conclusion that "groups studied in the literature are quite coarse." The MHTS was targeted on individuals with severe mental impairments, psychotic and affective disorders, whereas the SED is being tried on individuals who allege a much broader array of mental and general medical impairments. The more focused MHTS found differences in impact of IPS and integrated behavioral health treatments on younger beneficiaries with schizophrenia and on older beneficiaries who experienced depression. Older beneficiaries, presumed to have more work experience, fared better in employment outcomes. We will have to wait several years for the results of the impact analysis of the SED on different groups, as that study is not yet completed.

Although coarsely defined in the literature, some groups have more employment potential than others. The chapter points out, however, "even those SSDI beneficiaries with employment potential can face substantial labor market barriers and be at risk of financial hardship absent benefits." In my experience, this is true for individuals with disability due to mental impairments, who face hiring impediments based on prejudicial attitudes toward individuals with mental impairment. The MHTS and SED both illustrate the problems faced by such individuals in obtaining employment.

POINT OF DISAGREEMENT

I want to take exception to the characterization of individuals with mental impairments, particularly younger individuals, as having less severe impairments and thus having higher employment potential. Although this may be true of the findings of several studies reviewed in the chapter, this characterization is not uniformly true. Mental impairments often are invisible, and they tend to wax and wane. An individual with a mental impairment might seem to be able to work one day, and be unable to work on another, establishing a pattern of inconsistent work attendance and lack of productivity that is not conducive to full-time employment. The functional limitations imposed by mental impairments affect the full range of work demands, making any kind of work on a sustained basis a challenge. Younger individuals may have more years of potential for employment, which is a hopeful perspective, but many of the most severe mental disorders have their onset in the late teens and early twenties, and those individuals with earlier onset often are more severely functionally impaired than those with later presentations of their conditions.

The MHTS demonstrated the mixed experience of employment potential for individuals with mental impairments, particularly younger individuals. Although some 60 percent of individuals in the treatment arm of the MHTS had some level of competitive employment, compared with the control group at 40 percent, none stopped

receiving SSA disability benefits. Full-time employment was an elusive goal for this group, which sets up a final comment.

A GENERAL POLICY LESSON FROM THE MHTS AND SED

As noted in the chapter and in my comment above, none of the participants on the MHTS worked above the Substantial Gainful Activity (SGA) threshold of the SSA statutory definition of disability, and none exited SSA's disability programs. Based on my interviews, as a part of the SED evaluation team, I can report that beneficiaries and service providers, alike, are coming to view being on benefits and working some, within the SSA rules, as a good outcome. Participants in both the MHTS and the SED are working, most at less than the SGA level, but they are enjoying some of the benefits of social inclusion and work force participation. They do not achieve the policy goal of savings to SSA, but their mental health has improved, and they are experiencing a higher degree of social integration. SSA is to be applauded for supporting these demonstrations and these broader outcomes. And perhaps we should not be surprised that individuals found disabled by SSA are not able to work above SGA levels, even with substantial supports, because they have been found disabled under a very strict standard of disability.

Chapter 7

Comment

Nick Hart
The Data Foundation

More than 40 years ago, Lee Cronbach and Associates (1980) wrote that context matters in evaluation. Cronbach's plea for increased consideration of external validity in design and execution of evaluations is a poignant message for evaluators in the 21st Century, including with regards to Social Security Disability Insurance (SSDI) demonstration projects. Shortly after being sworn into office on January 20, President Joe Biden (2021) signed a first Executive Order directing agencies to reimagine how to assess and solve for inequities in government programs. Given this context about the appeals from the evaluation community for 40 years and new policy expectations to understand disparities and inequities, it is striking just how tremendous the gaps are in analyzing key impacts across relevant subgroups in past SSDI demos.

In "An Overview of Current Results and New Methods for Estimating Heterogenous Program Impacts" by Till von Wachter, the author's review of past demonstration projects alongside the strong evaluation infrastructure in place at the Social Security Administration (SSA) leads to a clear conclusion: analysis of beneficiary subgroups can be substantially improved. Some subgroup analysis did occur, as von Wachter notes, so this is not to say subgroup analysis was nonexistent at SSA. Yet only two of seven randomized evaluations analyzed differences by race and ethnicity, for example. The policy implications are vast and the risks for decisionmakers from gaps in knowledge where disparities or inequities may exist are tremendous, particularly when compared to the relatively low cost of increasing sample sizes or adjusting contracts for additional subgroup analytics in a program with $140 billion in benefit payments per year.

THE POLICY CASE FOR ANALYZING HETEROGENOUS IMPACTS

In 2015 when SSDI reforms were considered and negotiated by Congress and the White House, the policymakers turned to existing demonstrations to consider areas for reform, savings, improvement, and further study (McCann and Hart 2019). We know the policymaking community relies on insights from demonstration projects to inform actual policy—this is not theory; this is the practice and tradition for SSDI. In many respects, the historical use of evidence from SSDI demos is also a testament to the quality and effort from SSA and its partners in testing strategies for improving programs and services. Policymakers need relevant insights about subgroups that differentially experience SSDI as well as respond with variation across theorized program improvements. A single average treatment effect for a study population or sample is insufficient.

The major challenge that von Wachter identifies and articulates about the gaps in some demonstration projects of analysis of heterogenous effects is a theme that should have been addressed long ago as part of the design and planning of demonstration projects. In 2021 and beyond, improved analysis that bolsters external validity is an imperative. President Biden's Executive Order is an impetus; so too is the Foundations for Evidence-Based Policymaking Act (Evidence Act) and the ensuing evaluation standards that call for ethical evaluations in government that address "contextual factors that could influence the findings or their use" (Vought 2020, 5). The policy case for analyzing subgroups can be summarized to say there is both an expectation and an ethical obligation for the SSA and its evaluators to conduct analyses that explore heterogenous impacts across relevant subgroups. Doing so supports efforts to understand inequities and enables improved targeting of efficacious interventions to the individuals most likely to realize benefits.

IMPROVING ANALYSIS OF HETEROGENOUS IMPACTS RETROSPECTIVELY AND PROSPECTIVELY

The suggestions offered by von Wachter in his chapter for embedding heterogenous effects are practical, salient, low cost, and necessary. If anything, the major critique of von Wachter's chapter is that it simply does not go far enough in suggesting, given vitality of the topic, that much more should be and must be done by SSA in the future to encourage more subgroup analysis. These actions cannot solely be the responsibility of the research and evaluation community, but must be reflected by SSA leadership and stakeholders.

First, SSA should continue to plan for future evaluations to address heterogenous impacts when possible, with actions that could include these:

- ***SSA Evaluation Policy.*** In implementing SSA's published evaluation policy required by the Evidence Act and incorporating the required evaluation principles, SSA can explicitly reflect its own policy statement to prioritize particular types of analytics necessary to improve programs for SSDI beneficiaries as a key way to tailor findings to meet the needs of evaluation users (SSA 2020f).
- ***SSA Learning Agenda and Equity Assessment.*** In complying with the Evidence Act's requirements and the President's Executive Order, SSA can continue to effectively collaborate with its stakeholder communities and policymakers to identify key questions or themes to incorporate in future research and evaluation plans, to specifically study and address inequities or disparities. For example, key questions could include better assessment of perceptions and burdens, challenges in access, or denials of benefits by subgroups, including stratification of results by race and ethnicity when appropriate.

- ***Outcome Standardization.*** In leveraging the Evidence Act, SSA can begin to identify shared definitions and standards for interpreting particular outcome measures to enable improved comparability across studies, evaluations, and demonstration projects.

In addition to planning for the future, SSA can also potentially supplement insights from completed projects by leveraging its own existing data infrastructure to reanalyze past interventions through data linkages or sharing with other agencies and partners. The US Commission on Evidence-Based Policymaking unanimously recommended to the President and Congress in 2017 that federal agencies enhance data sharing and linkage capabilities (CEP 2017). SSA has a compelling case to be on the forefront of these linkage activities moving forward, including to apply current data to past research in order to generate new insights about long-term impacts across subgroups. Related, access to SSA data for retrospective evaluation through de-identified data sets or secure data enclaves could better support SSA's and the disability community's long-term evidence-building needs. Though some such activities may be underway at SSA today, updating procedures, regulations, and notices under the Privacy Act and SSA's authorizing statutes can be time intensive and burdensome, so prioritizing these activities should be a priority for SSA's chief data officer, evaluation officer, and other senior leaders.

CONCLUSION

SSA has a vital role in providing decisionmakers and stakeholders with relevant information about what works best, in what contexts, and for whom. Analysis of group effects in the future is essential for SSDI demonstration projects, and SSA should act upon the insights offered by von Wachter, including exploring new innovative mechanisms and approaches for addressing contemporaneous methodological and resource constraints. SSA must also reimagine its data and evaluation capabilities to ensure appropriate information is available as open data and for other evaluative activities, as well. SSA has historically been a leading federal agency for enabling evidence-informed decisionmaking, but there remains much room for progress and improvement in the years ahead.

Chapter 8

Benefits Counseling and Case Management

Vidya Sundar
University of New Hampshire, Occupational Therapy Department

The Social Security Administration (SSA) provides income support for older adults, individuals with disabilities, and families with low incomes through the Supplemental Security Income (SSI) and Social Security Disability Insurance (SSDI) programs. SSI is a means-tested program that is available to older adults, working-age adults with disabilities, and children with disabilities based on eligibility criteria related to disability, income, and assets. SSDI is a social insurance program and provides cash benefits to workers with disabilities and certain members of their family.[1] Both SSI and SSDI also offer entitlement to health insurance. SSI confers Medicaid eligibility for recipients, and SSDI beneficiaries become entitled to Medicare after receiving SSDI for 24 months.

These income support and safety net programs are an essential lifeline for millions of Americans who are unable to work and maintain economic self-sufficiency. For example, in 2019, SSI provided more than $52 billion in income support for 6.9 million individuals with disabilities (SSA 2019c); SSDI provided more than $11 billion in income support to 9.2 million working-age adults in 2018 (SSA 2019d). For individuals who are entering or re-entering the workforce, case management and benefits counseling services can assist in navigating the complex landscape of programs and policies that support work activity. This chapter will examine the impact of benefits counseling and case management services offered in the context of SSA demonstrations.

HISTORY, POLICY SETTING, AND CURRENT PROGRAM RULES

The Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act) was established to remove barriers to employment and to provide health care and employment services to SSDI beneficiaries and SSI recipients. The legislation recognized the need for benefits planning and assistance as a core service needed by individuals with disabilities who received SSI and/or SSDI. SSA established the Benefits Planning, Assistance, and Outreach (BPAO) program subsequent to the Ticket Act by entering into 116 cooperative agreements with community organizations across the nation (Livermore and Prenovitz 2010). By the end of 2001, all states had at least one entity that received funding from SSA to implement a BPAO program.

[1] Title II of the Social Security Act provides cash payments through SSDI to individuals who are younger than age 65, have earned sufficient work credits, and meet the definition of disabled.

The BPAO system was designed to assist SSI recipients and SSDI beneficiaries in maneuvering a complex set of public benefits programs, as well as to minimize disincentives and barriers to preparing for, retaining, or advancing in employment. Benefits specialists received intensive training on work incentives programs and provided services to individuals in person or over the phone. Under the BPAO, benefits specialists were instructed not to direct or influence beneficiaries and recipients regarding their employment-related decisions. Rather, BPAO counselors focused their services on education and sharing of accurate information in one or two sessions (Livermore and Prenovitz 2010; O'Day et al. 2009). In general, BPAO had mixed results in supporting the goals of the Ticket to Work program, which is to assist SSDI beneficiaries and SSI recipients in their transition to long-term employment and reduce their reliance on benefits. Findings from customer satisfaction surveys (Bruyere et al. 2007) suggest that BPAO was successful in providing accurate information to beneficiaries and recipients. However, findings from the State Partnership Initiative (SPI) suggested that benefits counseling may reduce earnings (O'Day et al. 2009). Subsequently, SSA determined that a greater emphasis on employment and in-depth services was needed to achieve the program's goals (O'Day et al. 2009).

In 2006, SSA's program priorities shifted from providing basic information about work incentives to providing long-term employment supports coupled with case management (O'Day et al. 2009). The Work Incentives Planning and Assistance (WIPA) program grew out of the BPAO program and was established in 2006 with the goal of increasing community partnerships, with a renewed focus on achieving employment outcomes. SSA recognized that SSDI beneficiaries and SSI recipients needed intensive services (rather than one or two the sessions that was typical in BPAO) to fully understand and use work incentives (O'Day et al. 2009). The overarching purpose of WIPA is to provide accurate information and counseling about the impact of work-related income on benefits and supplemental income programs. WIPA programs deliver services in four broad categories: work incentives planning; work incentives assistance; work incentives education, marketing, and recruitment of beneficiaries and recipients; and outreach services (O'Day et al. 2009).

WIPA is implemented through community work incentives coordinators (CWICs) whose role is to provide ongoing, comprehensive work incentives monitoring and management and to help SSDI beneficiaries and SSI recipients develop long-term work plans. CWICs provide both information and referrals and more intensive counseling services about benefits and employment. CWICs provide information tailored to beneficiaries' and recipients' needs and employment goals including any health insurance protections and work incentives that beneficiaries and recipients could qualify for. CWICs also verify eligibility requirements and educate beneficiaries and recipients about requirements to report wages and other income or change in work activity, thus helping them navigate a complex system of supports and services.

Eligibility for WIPA services is based on age (14 and older) and receipt of SSI, SSDI, disabled widower benefits, childhood disability benefits, or Medicare coverage based on disability status. WIPA services are prioritized for SSDI beneficiaries and SSI recipients who are working full-time or part-time, in the process of interviewing for work, or US military veterans who are working or seeking employment. WIPA services are also available to transition-aged youth (ages 14–24) and US veterans who are considering working.

WIPA counselors can be reached via a referral from a help line or by contacting WIPA offices directly. Once contact has been established, WIPA counselors work with SSDI beneficiaries and SSI recipients to gather information on current benefits and goals for employment. The CWIC verifies current benefits and over several sessions provides education and counseling on how work income could affect federal and state benefits, health insurance, and work supports. A written Benefits Summary and Analysis (BS&A) is provided to beneficiaries and recipients summarizing their current benefits and future goals for employment. Counselors who provide WIPA services are trained and certified through an SSA-funded Technical Assistance center. WIPA programs operate in close collaboration with several other programs and agencies, such as Ticket to Work, the Protection and Advocacy to Beneficiaries of Social Security (PABSS) grant program, Employment Networks, and Vocational Rehabilitation (VR) agencies.

WIPA programs facilitate the use of several work incentives such as impairment-related work expenses (IRWEs), Plan to Achieve Self-Support, Trial Work Period, and so on. IRWEs allow SSDI beneficiaries and SSI recipients to deduct the cost of certain impairment-related expenses from their earnings. PASS allows SSI recipients to set aside income and resources that will help them achieve self-sufficiency with the amount set aside not counting toward determining SSI eligibility or payments. Trial Work Period allows SSDI beneficiaries at least nine months to test their ability to work. During the Trial Work Period, SSDI beneficiaries will continue receiving benefits regardless of their income as long as work activity is being reported (SSA 2020e).[2]

It should be noted that SSA has generally not provided case management services to SSDI beneficiaries and SSI recipients outside of demonstration programs. However, SSA has included case management in many of its demonstrations, such as the Mental Health Treatment Study (MHTS) and SPI. In these demonstrations, the overarching purpose of case management was to provide information and referral to vocational assessments, employment services, and if needed, work incentives planning. Additionally, case management may be provided by Employment Networks or Vocational Rehabilitation agencies, funded or contracted by SSA.

The following section describes the theoretical frameworks for understanding benefits counseling and case management, followed by a review of SSA

[2] A more detailed description of all work incentives is available in the *Red Book* (SSA 2020e).

demonstrations and empirical research on benefits counseling and case management. Last is a summary of knowledge gained and research/policy recommendations for SSA.

THEORY AND IMPLICATIONS FROM THEORY

SSDI beneficiaries and SSI recipients experience systemic, structural, and personal barriers in seeking and retaining employment. Income support programs such as SSI and SSDI support a beneficiary's and recipient's ability to meet basic needs. However, the process of applying to and getting approved for SSI or SSDI benefits can be a long and arduous one. Because the approval process for SSI or SSDI requires demonstrated inability to work, some beneficiaries and recipients internalize this message and assume that they are unable to return to work even as their underlying condition stabilizes or improves (Miller and O'Mara 2003; Peikes et al. 2005). Yet other beneficiaries and recipients could desire to return to work but might not fully comprehend how their return affects their income, disability, and health benefits. The broad goal of *benefits counseling* is to provide information and counseling support so that SSDI beneficiaries and SSI recipients reach their employment goals and increase their economic self-sufficiency. Benefits counseling unfolds through the process of assessing and understanding the beneficiary's and recipient's employment goals, identifying viable options, sharing accurate information, and tracking and managing benefits (Delin, Hartman, and Sell 2012).

Benefits counseling can address the employment gap by providing in-depth analysis of pros and cons, step-by-step guidance, and follow-up monitoring of how well SSDI beneficiaries and SSI recipients understand and use the current programs offered by SSA. *Case management* involves collaborative assessment, planning, and mobilization of resources and care coordination. Within the context of SSA programs, case management is broader in scope than benefits counseling and can involve connecting beneficiaries and recipients with employment, housing, health care, and financial literacy resources.

Framework for Understanding Benefits Counseling and Case Management

Kregel and O'Mara (2011) describe four stages along an "employment continuum" that SSDI beneficiaries and SSI recipients go through while seeking employment. The first, *contemplative stage* is when beneficiaries and recipients are thinking about working but they generally lack any concrete vocational goals. In the second, *preparatory stage* beneficiaries and recipients have made an active choice to pursue employment goals and may have taken steps to work toward these goals. In the third, *job search stage* beneficiaries and recipients solidify their efforts by seeking employment support services, applying for jobs, interviewing, and so on. In the final, *employment stage* beneficiaries and recipients are successfully employed. Though some SSDI beneficiaries and SSI recipients remain in this last stage for a prolonged

period, others can experience challenges to sustaining work and consider leaving their jobs or scaling back. Kregel and O'Mara's conceptualization of an employment continuum closely aligns with the transtheoretical (or stages of change) model (DiClemente at al. 1991) that describes a cyclical process individuals go through when engaging in a new behavior.

Benefits Counseling

Golden et al. (2005) define benefits counseling as

> a set of benefits counseling strategies, services and supports that seek to promote work preparation, attachment, and advancement focusing on the enhancement of self-sufficiency and independence of Social Security Administration beneficiaries and recipients with disabilities through informed choice, which may result in decreased reliance on public benefit programs and increased financial well-being. (xvi)

The process of benefits counseling begins with the beneficiary or recipient seeking services. The counselor gathers information about the beneficiary's or recipient's goals, current benefits, and work situation. The counselor verifies the benefits and provides referrals to programs that may support the beneficiary's or recipient's work attempt or financial situation. The counselor educates the beneficiary or recipient about the effect of earnings on benefits, documents the counseling, and provides follow-up services as needed.

Benefits counseling programs were developed within the context of income support programs such as SSI or Temporary Assistance for Needy Families specifically to provide accurate information about complex benefits and work incentives to vulnerable populations that depend on them. Because they were developed for pragmatic reasons, the theoretical or conceptual foundation of benefits counseling programs is unclear.[3] Nevertheless, benefits counseling programs offered by SSA are somewhat aligned with well-established principles of employment or career counseling—such as creating a therapeutic alliance, being person centered, and the like. Two theoretical frameworks that can be used to understand and evaluate benefits counseling programs are *cognitive information processing theory* and the *solutions-focused approach*. Cognitive information processing theory suggests that making career and employment choices involves knowledge (understanding information) and feelings (self-awareness, motivation). One common aspect of benefits counseling and cognitive information processing theory is that both emphasize career readiness and the importance of case management services to assist an individual to attain their employment goals (Sampson et al. 2004).

---

[3] Golden et al. (2000) retrospectively proposed a theoretical framework for benefits planning and advisement after the Ticket to Work program was established.

Bezanson (2004) described a solutions-focused approach to employment and benefits counseling. Counselors using a solutions-focused approach help their clients develop an alternate vision for their future; one that allows them to acknowledge their problems and not circumvent the same. It is a goal-directed, future-oriented approach where the goal is to find solutions to problems rather than examine their causes (Trepper et al. 2006; Proudlock and Wellman 2011). Solutions-focused counselors take the role of an active listener and facilitator rather than an expert who is sharing their opinion. Through a series of open-ended questions, positive affirmations, and solutions-focused discussion, the counselor leads clients to uncover their motivation and develop realistic employment-related goals. Some commonalities between SSA benefits counseling approaches and the solutions-focused approach are the acknowledgement of reality (i.e., potential loss of benefits and health care) and direct action to address this potential loss by directing beneficiaries or recipient to other programs or employment to replace essential income and supports and improve beneficiaries' and recipients' economic position through work.

The solutions-focused approach is distinct from motivational interviewing or cognitive behavioral therapy. *Motivational interviewing* is a counseling practice that addresses ambivalent thinking and internal motivation to implement change in behavior. *Cognitive behavioral therapy* is a type of psychotherapy that helps individuals identify automatic negative thought processes that can influence their behavior and learn coping strategies to break away from the thought patterns. Although a solutions-focused approach has not been tested for its efficacy in benefits counseling, the model offers a framework to address barriers in a proactive manner.

Case Management

Case management is a complement to benefits counseling that integrates medical or social care services that address physical and social functioning with the goal of maximizing the individual's ability to recover and thrive in the community (Kanter 1989). The National Association of Social Workers (2013) defines case management as "a process to plan, seek, advocate for, and monitor services from different social services or health care organizations and staff on behalf of a client." In practical terms, case management is the mobilization, integration, and coordination of care in low-resource environments to maximize function (Ziguras and Stuart 2000).

Solomon (1992) described four distinct approaches to case management: assertive community treatment (ACT), strengths-based case management, rehabilitation case management, and generalist case management. ACT is a model provided in community settings rather than hospital or institutional settings. Clients have access to services at any time through on-call case managers, and the nature of services provided is individualized and intensive. The strengths-based case management has a strong theoretical foundation in positive psychology to leverage a person's strengths and

informal support networks to achieve desired outcomes.[4] Strengths-based approaches along with person-centered approaches, which are commonly used for youth case management, can especially be helpful in leveraging the strengths and motivation of beneficiaries and recipients to return to work. Rather than focusing on limitations, a strengths-based model focuses on capacities, skills, and abilities, regarding the individual as an active actor and co-director, rather than a passive recipient. Rehabilitation case management has the specific goal of service coordination among rehabilitation and medical professionals and case managers.

Regardless of the type of case management, some common denominators are flexibility, resourcefulness, creating structural supports, and building trust and rapport with clients.

REVIEW OF EMPIRICAL FINDINGS FROM SSA DEMONSTRATIONS

Benefits counseling and case management are critical components of SSA's demonstrations to help SSDI beneficiaries and SSI recipients navigate health care and employment supports. This section details specific SSA demonstrations that included substantial case management and benefits counseling components and analyzes outcomes attributed to those components.[5]

Benefit Offset National Demonstration (BOND)

BOND was created in response to a congressional mandate that SSA explore ways to increase the incentives for SSDI beneficiaries to return to work and subsequently decrease their reliance on SSDI benefits. BOND included two stages; Stage 1 tested the effect of a benefit offset for all beneficiaries; Stage 2 was implemented with a select group of volunteer and recruited beneficiaries to examine the impact of the offset and specific enhancements to counseling services. Beneficiaries were randomly assigned to one of three groups; (1) offset plus work incentives counseling (WIC); (2) offset plus enhanced work incentives counseling (EWIC); and (3) current-law rules, including benefits counseling (control). WIC was designed to be comparable to the WIPA services except that it was geared to address special provisions under BOND. EWIC included all services under WIC plus vocational skill and interest assessments, assistance, and support necessary for the beneficiaries to find and sustain employment. Findings discussed in this section are drawn from process and impact reports of BOND (Derr et al. 2015; Geyer et al. 2018; Gubits et al. 2018a/b).

Ten BOND sites were selected for the demonstration based on their geographic location, staffing, availability of employment services, and non-BOND benefits

---

[4] Positive psychology is the study of positive subjective experiences, emotions, traits, and strengths that enable individuals to thrive and flourish, unlike traditional psychology, which focuses on distorted thoughts and behaviors.

[5] SSA demonstrations that included benefits counseling and/or case management as a minor component are not discussed in this chapter.

counseling. BOND sites followed either a dispersed or dedicated staffing model for providing benefits counseling services. In the dispersed model, multiple staff devoted a portion of their time to provide BOND benefits counseling. In the dedicated model, all staff time was devoted to providing BOND counseling (Derr et al. 2015). The staffing model had implications for the nature of the counseling provided, as discussed at the end of this section. The following discussion will be limited to the impact of benefits counseling offered through Stage 2 of BOND.

As intended, there were major differences in the quantity and nature of counseling services provided to Stage 2 BOND beneficiaries. Treatment group 1 (WIC) and control group beneficiaries typically received information and referral services and basic information about work supports and incentives. Counselors for treatment group 2 (EWIC) were expected to proactively communicate with beneficiaries frequently, a process called "follow-up and follow-along services." WIC staff prepared written BS&A plans documenting how earnings may impact work incentives. EWIC staff developed BS&As, Employment Services Plans documenting barriers to employment, specific plans to overcome the same, and referrals to VR agencies or Employment Networks for additional evaluation and support (Gubits et al. 2018a/b). In general, beneficiaries in the EWIC group were more likely to have BS&As. Specifically, 65 percent of employed beneficiaries receiving EWIC had a BS&A, compared to 21 percent of beneficiaries in WIC. Similarly, beneficiaries who were looking for work and not in the labor force were more likely to have BS&As if they were in the EWIC group (Gubits et al. 2018a/b). EWIC counselors also reported spending a substantial amount of time on post-entitlement services such as completing SSA 820/821 forms, monitoring continuing disability review progress, and preparing Annual Earnings Estimates. In contrast, WIC counselors were required simply to respond to beneficiaries' inquiries (Derr et al. 2015).

There were also fundamental differences in how beneficiaries engaged with the WIC and EWIC counselors. It was typical for WIC counselors to provide a one-time information and referral service or to engage in brief contacts. Subsequently, the caseload for WIC counselors was much higher than EWIC counselors. On average, the EWIC caseload was about half the WIC caseload. As of January 2014, WIC caseloads per full-time-equivalent counselor ranged from 119 to 222 beneficiaries, whereas EWIC caseloads per full-time-equivalent ranged from 76 to 116 beneficiaries. Beneficiaries receiving EWIC were consistently referred for outside support and services. The largest number of referrals were seen among beneficiaries who were looking for work. More than half of the beneficiaries who already were employed when they joined the study ("at baseline") also received referral services, likely related to retaining or seeking different employment opportunities. As expected, once referrals were made, EWIC counselors followed up with the referral source to close any gaps in service delivery.

Ten performance benchmarks[6] for each BOND site were established prior to BOND implementation. The benchmark for initial contact and assessment was 100-90 percent; 80 percent for service coordination and pre-employment skills training; and 33 percent for WIC. It should be noted that performance reports for EWIC counselors were based on the number of engaged beneficiaries. All EWIC sites met performance benchmarks with one exception ("any contact last month"). EWIC sites well exceeded other benchmarks related to conducting needs assessments, skills assessments, service coordination, pre-employment skills training for those who needed it, and information and referral assessment.

In summary, there were considerable differences in the nature and impact of services provided through WIC and EWIC. These differences were compounded by extrinsic factors such as the caseload of counselors at each site and program, geographic factors, economic factors, and demonstration design. For example, because WIC enrollment (and the WIC caseload) was lower than expected, WIC staff were able to provide services that were more extensive than planned. This difference was more noticeable because EWIC enrollment exceeded expectations, thereby increasing the caseload for EWIC counselors. Geographic location of sites and staffing models could have also been confounding factors in determining the effectiveness of the services. Sites in rural locations likely had fewer employment-related services available near them, making it challenging for beneficiaries to receive essential support services (Derr et al. 2015).

Finally, the staffing model could have influenced the quality of services. In sites where a dispersed model was used, there was some anecdotal evidence of confusion between the different treatment options because staff in these sites provided BOND services infrequently. According to the findings from the Stage 2 early assessment report, several counselors noted that they were initially unfamiliar with how BOND offset worked and therefore were not able to provide accurate information to their clients (Derr et al. 2015). This lack on the part of BOND staff could have negatively affected program outcomes.

Preliminary evidence from BOND focus groups suggests that benefits counselors adapted to providing services over the telephone and that it was possible to maintain effective communication between counselors and beneficiaries that way. At the end of Stage 2, EWIC beneficiaries were more engaged with counselors, used more information and referral services, and interacted with their counselors more. Ultimately, there was no difference in earnings outcomes between the groups receiving WIC and EWIC services (Gubits et al. 2018a/b).

---

[6] Performance benchmarks established for engaged beneficiaries in the BOND EWIC group: any contact last month, barriers and needs assessment, skills assessment, Employment Services Plan, service coordination among those with documented need, pre-employment skills training, information and referral assessment, baseline assessment, BS&A, and Work Incentives Plan.

Promoting Opportunity Demonstration (POD)

POD began in 2018 and ended in June 2021 (Mamun et al. 2021). Its purpose was to address the complexities of work rules for the SSDI program by implementing a benefit offset paired with direct or indirect supports to facilitate the use of the offset. Eight states (Alabama, Connecticut, Vermont, and parts of California, Maryland, Michigan, Nebraska, and Texas) participated in the program. Beneficiaries who volunteered were randomized into two treatment groups and one control group. For beneficiaries in treatment groups, the Trial Work Period and the Grace Period were replaced by a set of new rules that included a benefit adjustment (reduction) of $1 for every $2 earned above the Trial Work Period threshold ($940 in 2021), called the POD "earning threshold" (treatment 1) or the "total monthly itemized IRWEs above the POD earning threshold" (treatment 2). In one treatment group, benefits were terminated after 12 months of $0 benefit; in the other treatment group benefits were not terminated during the demonstration. All participants in the two treatment groups received counseling on enrollment (Mamun et al. 2021). This aspect of POD was designed to address shortcomings of BOND by allowing eligibility for the benefit offset and assigning benefits counselors immediately upon enrollment (Wittenburg et al. 2021).

Findings from the interim evaluation report by Mamun et al. (2021) suggest that almost all treatment group members (more than 99 percent) received initial contact from their benefits counselors and less than half (38 percent) engaged in individualized work incentives counseling. In general, beneficiaries in the treatment groups were more work oriented (working or looking for work) than those in the control group. Although, beneficiaries reported that POD counselors were approachable and easy to work with, nearly half of treatment group beneficiaries indicated that the POD counseling services were not helpful for increasing their hours worked or earnings. Some beneficiaries reported that the information shared was not relevant to their situation because they were already working. Findings from the interim impact evaluation suggest the offset had no impact on earnings, SSDI benefit amount, or income. It is possible that the benefit offset did not provide a strong enough incentive for beneficiaries to change their work behavior (Mamun et al. 2021). A caveat in interpreting these findings is that the final evaluation report for POD was not available as of 2021, and the interim findings may change with additional data.

Promoting Readiness of Minors in SSI (PROMISE)

PROMISE was a joint venture of SSA with the US Departments of Education, Health and Human Services, and Labor. The demonstration was designed to address the systemic barriers faced by youth in meeting their long-term employment needs. The goal of PROMISE was to improve the provision and coordination of employment services anticipated to result in long-term economic self-sufficiency for the youth SSI recipient.

Five-year PROMISE demonstration grants were awarded in 2013 to five states and one consortium: Arkansas, California, Maryland, New York, Wisconsin, and a consortium of six western states known collectively as Achieving Success by Promoting Readiness for Education and Employment (ASPIRE). ASPIRE consisted of Arizona, Colorado, Montana, North Dakota, South Dakota, and Utah. The demonstration was extended an additional year in all states and ended in 2018 or 2019, depending on the project. The PROMISE final evaluation is anticipated in 2022. Findings discussed here are based on the interim services and impact report (Honeycutt, Wittenburg, Crane, et al. 2018), which focused on receipt of services after 18 months.

PROMISE had five major components: (1) strong intra-agency collaborations; (2) case management; (3) benefits counseling and financial education; (4) career and work-based experiences; and (5) parent training and information. Taken together, these five components were hypothesized to address individual, family, and institutional barriers to long-term economic self-sufficiency among youth. Youth SSI recipients ages 14-16 receiving SSI benefits and their families, residing in PROMISE service areas at the time of enrollment, were eligible to participate. Youth SSI recipients were randomly assigned to a treatment (PROMISE) or control (usual service) group. Between 2014 and 2019, each PROMISE site enrolled approximately 2,000 youth and their families (except California, where N=3,078).

Case management was the cornerstone of the PROMISE demonstration. Case managers played a central role in coordinating services and provided person-centered counseling, conducted needs assessment, and provided information and referral services. California PROMISE ("CaPROMISE") provided treatment group participants with the most extensive supports of all the sites; in addition to the five core components, California treatment group members received referrals for leadership and advocacy training, health and behavior management, access to assistive technology, and training in independent living. The following findings for PROMISE are drawn from Honeycutt et al. (2018), Levere et al. (2020), and Mamun et al. (2019).

Overall, the early outcomes for PROMISE participants were similar in all six PROMISE sites: members of the PROMISE treatment group demonstrated statistically significant positive outcomes after 18 months, compared with control group members. Each of the six programs increased the hours of transition services received, paid employment and support services, and family supports received (Levere et al. 2020).

In general, PROMISE programs offered case management services using one of three models: (1) in Arkansas and ASPIRE, case managers were employed by the lead agencies, and referrals were made to education, employment, and health-related services; (2) Maryland, New York, and Wisconsin hired their own case managers and supplemented additional community resources to support youth and their families; and (3) California offered services directly to participants and required their case managers to be certified in benefits counseling. It should be noted that at all sites any benefits

counseling provided to participants was provided by trained benefits counselors, but not all benefits counselors were case managers.

Case managers in all PROMISE sites met with youth and their families to provide benefits counseling coupled with financial literacy training. Financial training included budgeting, bank accounts, self-sufficiency, and consumer credit. The structure of financial training programs varied among sites. Some sites (Maryland, ASPIRE) started with contracted group training sessions and transitioned to individualized training after enrollment numbers were low for the group sessions. In addition to financial training, California, Wisconsin, and ASPIRE sites provided financial coaching and opportunities to increase savings through Individual Development Accounts, state-matched college savings plans, and Achieving a Better Life Experience (ABLE) accounts (Honeycutt et al. 2018).

Findings from the interim process and implementation analysis highlighted features of each PROMISE site that contributed to the outcomes (Mamun et al. 2019). In Arkansas, more than 92 percent of youth were engaged in PROMISE three years after the program began. Arkansas was able to accomplish this by converting some of its recruitment staff to retention staff and increasing outreach efforts. About 59 percent of youth received case management services; and almost all participating youth had identified career goals and plans to achieve the same. Two-thirds of participating youth had started summer work experiences, and about 25 percent completed the work experience for two summers. Parents in the Arkansas PROMISE were also highly engaged in the program. At the end of three years, about 87 percent of parents of participating youth had their own PROMISE goals and were referred to education and employment services (Mamun et al. 2019). At the end of 18 months, Arkansas PROMISE was also able to increase employment rate, hours worked, and earnings of treatment group youth in comparison with control group youth.[7] The program did not have any impact on youth education or self-determination outcomes or parent employment or earnings (Mamun et al. 2019).

ASPIRE prominently featured case management services. ASPIRE case managers were supposed to meet with all youth participants and their families for at least 30 minutes once a month to provide benefits counseling, financial education, information and advocacy support, and self-determination support. The interim process and impact analysis indicated that 86 percent of youth remained engaged in ASPIRE. However, the program fell short of its case management and benefits counseling goals; only 47 percent of all youth participants received case management services, and most case management contacts were less than 20 minutes; and 46 percent of families received benefits counseling services. ASPIRE sites met their goal for career engagement, where 31 percent of youth had engaged in competitive employment by the second year. The program also had a positive impact on the receipt

[7] For treatment group members, PROMISE increased employment by 31 percentage points, average hours worked by 2.7 percentage points, and average earnings by 162 percentage points compared to the control group (Mamun et al. 2019).

of transition services by youth and families, but no impact on youth education or self-determination outcomes, parent education, or earnings (Mamun et al. 2019).

CaPROMISE focused on providing intensive family-centered case management and work experiences for youth. CaPROMISE had a positive impact on youth employment and earnings and on parent earnings, education, and training in comparison with the control group.[8] The positive impacts on parent outcomes could be linked to the program's strong emphasis on family-centered services.

Maryland, New York, and Wisconsin experienced early challenges in implementation. In Maryland, the PROMISE program did not meet its benchmark of providing 8 to 10 case management contacts per month for youth and their families, possibly due to staff dedicating most of their time to recruitment rather than retention. In New York, case managers were responsible for providing intake evaluation and providing case management services or had additional non-PROMISE job duties that limited their availability to provide PROMISE services. In Wisconsin, there was a low uptake of services by youth and their families. Although 95 percent of families engaged in case management, only 65 percent were referred to job development, 39 percent had paid work experiences, 36 percent had any contact with a benefits counselor, and 14 percent completed soft skills training. A combination of factors such as poor referral rates, non-PROMISE-related demands on counselor time, and conflicting family priorities may have contributed to the low uptake (Mamun et al. 2019). Maryland PROMISE program did not meet its benchmark of providing 8 to 10 case management contacts per month for youth and their families, possibly due to staff dedicating most of their time to recruitment rather than retention. At the 18-month analysis, Maryland and New York were successful in increasing services delivered and youth earnings but did not have a detectable impact on other outcomes. Wisconsin also demonstrated increases on program participation, youth earnings, and youth health insurance coverage (Mamun et al. 2019).[9]

Further analysis of PROMISE services (Levere et al. 2020) suggest that youth and family services were associated with favorable outcomes. However, because youth and family services were provided concurrently, it is not possible to disentangle the impact of each separately and no causal inference can be made about the impact of each. Benefits counseling along with networking, support, parent training, and information on their youth's disability were bundled as "youth-oriented family

---

[8] CaPROMISE increased by 5 percentage points the share of parents reporting that they or their spouse had attended or completed job skills training or education during the 18 months following random assignment. CaPROMISE increased the self-reported earnings of parents in treatment group by $122 compared to the control group. However, a similar increase was not observed in SSA records, possibly due to differences in reference periods for data collection (Mamun et al. 2019).

[9] At the 18-month evaluation, Wisconsin PROMISE showed a 1 percentage point impact on health insurance coverage for youth in the treatment group compared with the control group—small but statistically significant (Mamun et al. 2019).

services." The "family-oriented family services" bundle included case management, education or training supports, employment-promoting services, and financial education services provided to family members other than the youth receiving SSI (Levere et al. 2020).

The use of youth-oriented family services had a moderate, non-significant association with youth outcomes after controlling for youth and family characteristics. Typically, youth who used services had better outcomes than youth who did not (except for SSI payments). However, there was no statistically significant relationship between use of family services and youth outcomes. Although the findings suggest association and do not demonstrate a causal relationship between either bundle of services and youth outcomes, they provide preliminary evidence of the potential importance of those services in the youth's transition process.

Nye-Lengerman et al. (2019) examined emerging lessons from the PROMISE demonstration. Their findings suggest that successful PROMISE programs demonstrated flexible service delivery models, strong leadership, solid interagency collaboration, opportunities for professional development for staff, and family engagement.

Youth Transition Demonstration (YTD)

The YTD was a set of projects aimed at youth who were receiving or at risk of receiving SSI benefits. For the evaluation, SSA selected six sites from a larger group of sites that had participated previously through cooperative agreements or as pilot programs. Three projects entered the evaluation in 2006-2007, and three in 2008. In total, these six projects randomly assigned more than 5,000 youth who volunteered to participate (Fraker, Mamun, et al. 2014).

Each site was able to define its specific target population and approach, with most serving SSI recipients and all evaluation sites offering a set of core services developed for YTD based on the *Guideposts to Success* model (NCWD/Y 2005, 2009). These included work-based experiences, system linkages, youth empowerment, family supports, social and health services, and benefits counseling. The intervention also involved waivers to SSA benefit rules that relaxed the conditions around the Student Earned Income Exclusion, the Plan to Achieve Self-Support, and Individual Development Accounts; increased the Student Earned Income Exclusion; and provided continued benefit payments and Medicaid coverage under Section 301 for the period of participation in YTD for those found no longer disabled or who turned 18 and did not meet the adult definition of disability (Rangarajan et al. 2009).

The model of benefits counseling and case management used in YTD was characterized by two features. First, they were integrated with a larger set of services and supports provided as a way to facilitate access and use of other services and supports, both within and beyond the program. For example, benefits counseling was tailored to explain both the waivers for which participants were eligible and the regular SSA rules that would apply after the program had ended (Rangarajan et al. 2009). The

second feature was substantial local flexibility, which reflected differing participant needs, service environments, and the capacities of and choices made by the organizations implementing YTD programs. For example, the Maryland site served youth who were not currently receiving SSI benefits, so benefits counseling emphasized other benefits, such as the Supplementary Nutrition Assistance Program (SNAP) and Temporary Assistance to for Needy Families (TANF) (Fraker, Baird, et al. 2012). The West Virginia site operated in a fragmented service environment, where most service providers had limited capacity for outreach and so depended on youth seeking out their services. For this reason, an important part of case management for the West Virginia program was helping youth and their families to identify supports (Fraker, Mamun, et al. 2012). The YTD program based in Bronx County (NY) structured many of its activities around "Saturday Sessions" in which youth and their families participated in group activities. Elements of benefits counseling and case management that could reasonably be provided in a group setting, such as general information on SSA benefit rules, were incorporated into these sessions, with additional case management and benefits counseling provided individually as needed (Fraker, Black, Broadus, et al. 2011).

The variation across demonstration sites allowed for the exploration of many different models, but also makes it difficult to aggregate findings across sites. Indeed, the final report considers each of the sites separately (Fraker, Mamun, et al. 2014). Also, as is the case in most demonstrations reviewed here, it is impossible to isolate the effects of case management and/or benefits counseling, as both were integrated into a larger program. Four of the six programs increased at least one measure of earnings and/or employment, and all but one increased at least one measure of youth income, often by increasing SSI benefits as extended eligibility through Section 301 (Fraker, Mamun, et al. 2014). However, it is unclear to what extent these results were caused by case management or benefits counseling.

State Partnership Initiative (SPI)

SPI was designed to respond to persistent employment issues, low rates of employment, low earnings, and inadequate use of work incentives programs by individuals with disabilities. SSA partnered with the US Department of Education's Rehabilitation Services Administration (RSA) to provide funding for this demonstration. Eighteen states participated in SPI between May 2001 and September 2004, of which 12 were funded by SSA and 6 by RSA. The focus of the initiatives varied slightly depending on the source of funding: SSA-funded states provided information, better access to vocational supports, and modified program rules (waivers) to allow for more earning and saving. RSA-funded states focused heavily on changing service delivery models. Participating states designed their own interventions, choosing from a menu of seven barriers to address that are most frequently faced by SSDI beneficiaries and SSI recipients (Peikes et al. 2005). All states provided benefits counseling; all states except North Carolina, Ohio, and

Oklahoma provided Medicaid waivers and buy-ins. Most states provided one or more employment services in the form of placement assistance or case management (Kregel 2006a).

Though each state differed in how it implemented benefits counseling, the common elements among the states were information and referral, problem solving, benefits assistance, benefits planning, and long-term benefits management. Three states (New Hampshire, New York, and Oklahoma) used random assignment to configure their treatment and control groups (Peikes et al. 2005). Each of these three states offered multiple intervention packages. New York offered two packages: The first package provided benefits counseling and tested changes to SSI regulations that allowed SSI recipients who worked to retain and save more money.[10] The second package added employment services to help participants find, apply for, and maintain employment. Oklahoma offered voucher services to participants who had a mental illness, received SSI, and were not employed at intake. The vouchers allowed SSI recipients to obtain vocational services from vendors of their choosing. All its participants received benefits counseling (averaging 10 hours per month) and job services through the vouchers (averaging 5 hours per month). More than three-quarters of participants received case management (averaging 7 hours per month) (Peikes et al. 2005). Supported employment, placement assistance, situational assessment, job training, psychosocial rehabilitation, job accommodations, or transportation assistance were offered less frequently. New Hampshire provided SSI recipients and SSDI beneficiaries with a choice of and control over their vocational services through the assistance of a service resource.

SPI benefits counseling interventions tended to produce modest impacts on employment and earnings. In New York, where benefits counseling was offered in conjunction with employment services, the package was found to be more effective than New Hampshire's intervention providing counseling only. New Hampshire saw a 30 percent decline in employment rates for the treatment group compared to the control group. Qualitative case report data from New Hampshire's SPI project indicate that a few participants chose to leave jobs to pursue education, training, and certification that would further their career goals (Cloutier et al. 2006). It is possible that the decrease in earnings could be attributed to this shift into education, but it is unknown how much of the drop in earnings can be attributed to this cause.

In Oklahoma, the intervention focused on individuals with psychiatric disabilities, who received benefits counseling, case management, and a voucher for employment services. Employment rates for treatment group members in New York and Oklahoma increased 9 to 18 percentage points. Earnings, on the other hand, did not change (Peikes et al. 2005).

---

[10] SPI demonstration tested waivers to SSI regulations that allowed recipients to retain more of their earnings and benefits counseling. In other. For a detailed description of the waivers, see Peikes et al. (2005, Appendix A).

There are several possible explanations for these findings. First, it is possible that benefits counseling in the absence of other employment support services is of little value. Benefits counseling coupled with vouchers for employment services, case management, and more important, assistance to find and keep a job could be effective. The combination of benefits counseling with employment services is particularly important, because the pattern of results suggests that simply providing information via benefits counseling without assistance with job search and placement will not affect employment status or earnings.

Second, about 79 percent of individuals who participated in this demonstration experienced mental or emotional disabilities, and 14 percent had physical disabilities (Kregel 2006b). It is possible that the intervention has a differential effect on different subpopulations.

Last, the follow-up period for the evaluation was likely too short to detect impacts that might take more time to emerge. Three months was probably not enough to capture any true changes in earnings that could have occurred due to benefits counseling. It is unlikely for any new employee to experience substantial increases in wages within their first three months. Longer-term follow-ups, as long as four years, might be necessary to capture true changes in earnings.

Accelerated Benefits (AB)

The AB demonstration was authorized by Congress in 1999 to examine alternatives to SSDI's 24-month waiting period for Medicare. The rationale behind AB was that SSDI beneficiaries could experience serious health care needs because of poor health and limited functioning. Acknowledging the relationship between health and employment, AB was designed as a five-year program to test the impact of providing health care services on overall health, employment outcomes, and reliance on SSDI benefits (Michalopoulos et al. 2011). Two versions of AB were tested; both versions provided health care benefits to SSDI beneficiaries until they were eligible for Medicare. The second version of AB, called AB Plus, offered additional services in the form of telephone counseling to help beneficiaries navigate the health care system and return to work if they desired to do so.

AB Plus participants were provided access to telephone counseling services through a health care management company (Weathers et al. 2010). Specifically, AB Plus participants received a baseline assessment and were assigned a nurse, coach, or both. Nurses assisted with navigating the participants' health care needs. Coaches, who were psychologists or social workers, guided participants through a Progressive Goal Attainment Program to reduce psychosocial barriers to rehabilitation progress, promote reintegration into life-role activities, increase quality of life, and facilitate return to work.[11] Its overarching goal was to encourage active steps toward seeking

[11] For information about the Progressive Goal Attainment Program (PGAP): http://www.pdp-pgap.com/pgap/en/index.html.

employment by optimizing work-life roles, by using behavioral coaching strategies to minimize barriers to rehabilitation. The final component of the AB Plus program was employment benefits counseling, which was available to participants who showed interest in returning to work.

Between October 2007 and January 2009, the demonstration enrolled 1,939 participants in a treatment group (AB or AB Plus) or the control group. Process and outcome evaluations of the AB demonstration were conducted. Both health-related outcomes (e.g., health care use, health status, unmet needs) and employment-related outcomes (e.g., job preparation, job search, use of work supports) were tracked (Michalopoulos et al. 2011).

Although the AB intervention did not cause changes in participants' labor market outcomes, the AB Plus intervention had a significant short-term impact on employment. Participants in AB Plus saw modest increases in short-term employment compared with the control group: a 4.6 percentage point difference in receipt of rehabilitation or employment services, 3.3 percentage point difference in receipt of services from the Ticket to Work program, and most notably, a 5.3 percentage point difference in employment during the second calendar year following enrollment. Subsequently, participants also demonstrated an increase in annual earnings of $831 by the second year (Weathers and Bailey 2014). In general, AB plus participants who used employment and benefits counseling had experienced higher levels of labor market activity. Weathers and Bailey 2014 note that "12.3 percent of employment and benefits counseling users participated in the Ticket to Work program, compared to 3.1 percent of those who did not use those services" (Weathers and Bailey 2014, 604).

However, these gains were short lived, not sustaining into the third year after enrollment in the study. It is possible that fear of losing their SSDI benefits triggered beneficiaries to adjust their labor market participation to preserve benefit receipt.

Subgroup analyses revealed that the earnings gain was highest in beneficiaries ages 45–49 or younger than age 40. Beneficiaries with a bachelor's degree and those experiencing respiratory and sensory limitations experienced higher gains in earnings than those without.

Findings from the AB demonstration suggest that providing a health insurance package (AB) is not sufficient to increase labor market activity. However, adding employment and benefits counseling (AB Plus) was marginally effective in improving short-term earnings in a small but substantial group of new beneficiaries.

Mental Health Treatment Study (MHTS)

The MHTS aimed to increase the employment outcomes (including earnings), health status, and quality of life of individuals with schizophrenia who were SSDI beneficiaries (Frey et al. 2011). It was designed on the heels of a large body of evidence on medical management integrated with supported employment services to improve employment outcomes of people with schizophrenia.

The study was fielded between November 2006 and July 2010 and targeted beneficiaries with schizophrenia or an affective disorder in 23 sites throughout the United States. Sites were eligible to participate if they had the capacity to deliver behavioral health interventions and had documented fidelity in delivering supported employment. Participants were eligible if they were ages 18–55, experiencing schizophrenia or affective disorders, and not experiencing any terminal illness. More than 2,200 beneficiaries were randomized into treatment and control groups and participated in the intervention for 24 months. The treatment group received supported employment services and evidence-based mental health services and supports including benefits counseling (where possible). The study design included strict and periodic quality management reviews conducted by nurse care coordinators. The primary outcome measures of interest were employment rate, earnings at main job, hours worked, number of months employed, health status, and quality of life. The analytical plan included both exploratory and confirmatory hypotheses, reflecting a well-planned analysis design that is present in some but not all demonstrations.

There was a 20 percentage point difference in the employment rates between beneficiaries in the treatment group and the control group. Beneficiaries in the treatment group were more likely to be employed in any job as well as in competitive jobs. There were also statistically significant differences in employment rate among subgroups based on age, gender, diagnosis, and educational status. Beneficiaries in the treatment group were also more likely to be steady workers rather than erratic or minimal workers. Factors that predicted employment rate included being enrolled in the treatment group, baseline physical health, previous work experience, and months receiving SSDI. A large, statistically significant difference was observed between the treatment group and the control group on earnings.

A closer examination of the benefits counseling and case management services delivered through Individual Placement and Support (IPS) revealed that 69 percent of participants in the treatment group received benefits counseling and 54 percent received mental health case management services at any time during the two-year study period. It should be noted that of the 54 percent who received the case management services, about a third used off-site locations. This lack of comprehensive onsite case management services, a central component of the IPS model, could have negatively influenced the program outcomes. Case management services also were not tracked uniformly across all sites because sites did not include case manager interactions in their monthly data collection form. Some sites provided case management services by telephone; however, these services were not compensated, leading to their possible deterioration or discontinuation.

Overall, evidence through the rigorous evaluation of MHTS and other empirical research (discussed later in this chapter) suggests that supported employment increases employment for beneficiaries (Frey et al. 2011). Because of the integrated nature of services provided through IPS's supported employment, it is challenging to isolate the impact of benefits counseling or case management only. Nevertheless, case

management, benefits counseling, and job placement services are critical components of supported employment programs,[12] which have demonstrated meaningful improvements in employment status.

Supported Employment Demonstration (SED)

The SED is a multi-component intervention targeting applicants for SSDI and SSI with mental impairments who were denied disability benefits on initial determination (Marrow et al. 2020; Taylor et al. 2020). SED is based on evidence-based supported employment and integrated behavioral health components. SED participants also receive additional funds to cover copays for medical treatment, work-related expenses, and other financial barriers (Marrow et al. 2020). The intervention aims to improve clinical recovery, increase employment, and subsequently keep individuals from needing SSDI or SSI.

SED was designed on the heels of MHTS, but the primary objective of SED is to test the effectiveness of supported employment services at an earlier stage. Participants in SED were randomly assigned to one of three treatment arms: Full-Service, Basic, and Control (approximate N=1,000 each arm). The multi-component intervention for its Full-Service and Basic treatment arms was delivered by a team of experts including a team lead, at least one IPS specialist, and a care manager. In addition, the Full-Service treatment included a nurse care coordinator. Benefits planning was embedded within the services provided through the IPS model of supported employment for the Full-Service and Basic treatment arms.

Findings from the interim process evaluation report from the first two years of the demonstration show that more than half of the sites (57 percent) were able to achieve high fidelity of implementation (Marrow et al. 2020). Participants experienced many unmet needs related to food, shelter, and medical care. SED staff had to leverage resources in the community to provide wraparound services to meet participants' medical and care needs. Overall, initial engagement with the SED team was positive (more than 92 percent) and 40 to 50 percent of participants continued to meet with their SED specialist monthly (Marrow et al. 2020). Final data on employment and clinical recovery outcomes are due in 2022.

[12] Traditional supported employment programs may include services such as career exploration, job search, customizing job duties or work schedules. In contrast, the IPS model promotes recovery through work and is defined by the following principles: (1) a focus on competitive employment, (2) rapid job search, (3) eligibility based on client choice, (4) attention to client's preferences in employment services and supports, (5) integration of employment and clinical services, (6) time-unlimited support, and (7) systematic job and employer relationship development. Some supported employment programs may incorporate other services such as cognitive behavioral therapy, occupational therapy, etc.

Project NetWork

Project NetWork was fielded between 1991 and 1995 to test the impact of case management as a means of promoting employment among persons with disabilities (Kornfeld and Rupp 2000). The demonstration targeted both SSI recipients and SSDI beneficiaries. Participants were recruited via two streams: (1) SSI applicants who volunteered and (2) SSI recipients and SSDI beneficiaries who were recruited through an outreach effort. Study participants were assigned to either a treatment or a control group. Those in the treatment group received case management, benefits counseling, and individualized employment services. Under the Project NetWork waiver, program rules that were considered a disincentive to working were waived. For SSDI beneficiaries, the Trial Work Period was suspended for the first 12 months. In other words, months with earnings did not count against the Trial Work Period and did not result in benefit suspension. For SSI recipients, Project NetWork waivers prevented a continuing disability review from being triggered (Kornfeld and Rupp 2000; Rupp, Bell, McManus 1993).

Project NetWork was implemented in eight sites and tested four case management models. In the first three models, there were differences in the organizational role of the case manager. In the first intervention, case management was provided by SSA staff; in the second, case management was provided by contracted rehabilitation organizations; in the third, case managers from Vocational Rehabilitation agencies were "outstationed" in SSA offices. The fourth model was designed to be less intense and focus on information and referral services, rather than direct services to clients (Kornfeld and Rupp 2000).

Findings from the demonstration revealed that participants in the intervention groups received more return-to-work services than the control group did, including benefits counseling, physical therapy, work assessments, and job search services. There was a statistically significant increase in earnings for the treatment group compared to the control group for the first two years following random assignment. However, those differences did not sustain during the third year following random assignment.

SSI and SSDI benefit receipt did not change between treatment and control groups during the follow-up period. Similarly, self-rated health also did not change. Further analysis of earnings and receipt of benefits by impairment subgroup did not reveal any major trends except for beneficiaries who experienced musculoskeletal disorders. Beneficiaries with musculoskeletal disorders saw a 2.1 percent reduction in receipt of SSDI benefits. Despite modest gains observed, the cost of administering Project NetWork exceeded the benefits realized from the program.

Project NetWork findings suggest that case management might not be relevant for all SSDI beneficiaries and SSI recipients, given that there was no sustained difference in earnings between the treatment group receiving case management and the control group receiving information and referral. Case management services might be of value

when beneficiaries or recipients are less job ready or experience limitations that they require coordination of vocational, rehabilitation, and employment services.

EVIDENCE FROM ADDITIONAL EMPIRICAL RESEARCH

Empirical research outside of SSA demonstrations is scant on the intersection of benefits counseling or case management and welfare programs. A few studies have examined the effectiveness of the case management approach using administrative data. The following section describes research conducted with the goal of demonstrating the effect of SSA programs, with subpopulations or using methodology that addresses some of the shortcomings of SSA demonstration designs.

Braitman et al. (1995) examined the barriers experienced by employed and unemployed clients in a case management program. Although the authors initially hypothesized fear of losing benefits, lack of family support, and transportation as primary barriers, their findings suggest that personal factors such as motivation, ability to tolerate criticism, and ability to self-initiate were ranked as important factors that determine employment. Many participants, regardless of employment status, rated illness-related symptoms as a barrier. Case managers need to be aware of the debilitating effects of illness, its side effects, and how that might affect work performance.

Bloom, Hill, and Riccio (2003) used consolidated data from multiple welfare programs to demonstrate the value of case management. Personalized attention in the form of spending time to understand the complex life circumstances of clients and their families and tailoring services to their specific needs was considered a critical component of a successful case management program. Peck and Scott (2005) examined the use of a Case Management Screening Guide to improve ability of case managers to identify the unique needs of their clientele. Use of the screening tool was associated with increased understanding of the strengths and weaknesses of clients, the number of employment services used by clients, case closures, and work-related activity. However, use of the screening tool had no impact on five-year employment status.

Evidence supporting the use of a case management model to improve employment status is strong within certain subpopulations such as individuals with mental health issues. A strengths-based case management model that was implemented with high fidelity had increased competitive employment of participants after 18 months of intervention (Fukui et al. 2012). A high-fidelity model of case management is characterized by structural components (low caseload sizes, periodic group supervision including case presentations, etc.) and practice components (use of the strengths assessment and recovery plan tools, use of natural supports in the workplace, and in-person service delivery) (Fukui et al. 2012).

Evidence-based practices such as assertive community treatment (ACT) and supported employment incorporate case management as a critical component. Both interventions are frequently used with adults with mental health issues. ACT integrates

principles of traditional rehabilitation and case management into one program. The key hallmark of ACT is the provision of case management and rehabilitation through one integrated team, where the case managers broker services and provide information and referral, and the rehabilitation team addresses function and employment-related goals. ACT case managers are also characterized by smaller caseloads of approximately 30 clients each and a well-defined job description (Boyer and Bond 1999; Ellison et al. 1995). Future adaptations of case management programs for use by SSDI beneficiaries or SSI recipients with mental health issues could leverage ACT best practices such as deploying integrated teams that address medical, social, and employment related issues.

Olney and Lyle (2011) conducted in-depth interviews with 12 SSDI beneficiaries and SSI recipients to understand the employment barriers they experienced. Findings suggest that participants were leery of losing the safety net of benefits. Some participants intentionally kept their earnings low, did not pursue career advancement opportunities, and sought out low-paying jobs. Participants engaged in cost-benefit analysis before deciding when, where, and how much to work. Those who were supported by family members' health insurance plans were more likely to reduce their reliance on SSA benefits.

The timing of benefits counseling is an important determinant of employment outcomes. In addition to traditional services, VR agencies that provided timely[13] benefits counseling observed greater SGA-level employment compared to agencies that waited to provide services (Honeycutt and Stapleton 2013). A similar effect was observed in the Kentucky Substantial Gainful Activity demonstration sponsored by the US Department of Education (Martin and Sevak 2020), adding further evidence that providing benefits counseling early was a critical component of successful programs. Martin and Sevak (2020) noted that eligibility determination for participants in the Kentucky SGA demonstration was completed within 2–10 days of initial contact, team meetings were conducted within 30 days from initial contact, and Individualized Plan of Employment goals were established within 30–60 days from initial contact.

Evidence supporting the value of a written benefits analysis plan is mixed. A written benefits analysis plan was not associated with increased earnings and had a modest impact on employment status for those employed for at least one quarter (Wilhelm and McCormick 2013). For transition-age youth who received benefits counseling and employment services (through PROMISE), the provision of benefits

[13] Timeliness of services was measured as Usual Wait Time. More than half of the study sample in Honeycutt and Stapleton (2013) had wait times of three months or less and about 90 percent had wait times of nine months or less. Of course, there are individual, agency-level, and state-level variations in timeliness of services. In general, each additional month of waiting for services is associated with a 1.2 percentage point reduction in SGA months (months of earnings at or above the SGA level after VR application, as recorded in SSA's Disability Control File).

counseling preceded by a "warm handoff"[14] steeped in trust and client-centered practices may be helpful (Schlegelmilch et al. 2019). Rather than being provided detailed written summaries of benefits analysis, youth and their families appreciated being met "where they are" with limited, relevant, and bite-sized information that was not overwhelming.

In general, demonstration projects did not control for or consider the effect of non-random selection of participants. When participants are included in a study on a volunteer basis, they can differ from the broader sample pool of SSI recipients and SSDI beneficiaries in many ways. Volunteer participants could be more motivated to work on their employment goals or have easy access to employment services within their community. It is likely that these differences in sample characteristics contributed to or caused changes in employment status and earnings, rather than the actual intervention provided. To address the issue of non-representative sample selection, Nazarov (2013)[15] and Iwanaga et al. (2021)[16] used data on SSI recipients and SSDI beneficiaries and employed quasi-experimental methods and propensity score matching to examine the effects of employment and benefits counseling services on earnings, hours worked, and labor market activity among adults and youth, respectively.

Findings suggest non-significant differences in case closure (due to employment) between those enrolled in benefits counseling and the control group. Findings from both studies suggest statistically significant increases in the estimate for earnings and hours worked among adults and young adults after controlling for non-random selection. Nazarov (2013) observed a 17 percent increase in earnings and a 20 percent increase in hours worked for adult beneficiaries and recipients in the treatment group. Iwanaga et al. (2021) noted that the youth in the treatment group worked fewer hours but had higher earnings than the control group. Taken together, these studies demonstrate the importance of controlling for non-random selection to uncover the *impact* of the intervention.

Tremblay et al. (2004, 2006) used a quasi-experimental design with two groups of matched comparison groups to examine the impact of specialized benefits counseling among participants enrolled in Vermont SPI. Five variables were used to

---

[14] When families seemed reluctant to transition from a VR counselor to a benefits counselor, the process of instilling trust and facilitating a rapport was described as a "warm handoff" (Schlegelmilch et al. 2019).

[15] Nazarov (2013) used data from the Case Management Administration System from the New York State Adult Career and Continuing Education Services (ACCES-VR). Study participants (N=38,125) were SSI/SSDI beneficiaries who received VR services between October 2003 and October 2009 and who had fully developed Individualized Plans of Employment.

[16] Iwanaga et al. (2013) used data (N=19,383) from the Case Service Report (RSA-911) for the 2018. The inclusion criteria for this study were (1) ages 18–35 (i.e., transition-age youth and young adults), (2) a primary diagnosis of intellectual disabilities at intake, (3) SSI recipients at intake, and (4) received VR services.

draw two comparison groups:[17] (1) experience as a VR consumer, (2) experience as an SSDI beneficiary or SSI recipient, (3) primary VR disability, (4) start date for VR services, and (5) time elapsed between eligibility and initiation of VR services. These variables were previously demonstrated as having an impact on employment outcomes. Comparing earnings over four years, the group that received specialized benefits counseling fared consistently better than the comparison groups. The adjusted difference in earnings between the intervention group and comparison groups was more than $1,200 per person per year. Analysis of within-group differences for the counseling intervention group indicated an almost $500 increase in earnings by the seventh and eighth quarters from baseline.

These findings add to the growing body of evidence supporting the effectiveness of benefits counseling programs, even after controlling for race, gender, disability type, and Social Security beneficiary type (Tremblay et al. 2004; Tremblay et al. 2006).

KNOWLEDGE GAPS AND LESSONS FOR POLICY FROM DEMONSTRATION PROGRAM FINDINGS

Findings from SSA demonstrations contribute to the growing body of evidence on the effectiveness of employment and benefits counseling. There is moderate to weak evidence that case management and benefits counseling contribute to increases in employment or earnings or to decreases in reliance on SSI or SSDI benefits. These findings should be interpreted in the context of programmatic, structural, and contextual differences among the demonstrations. The following section summarizes lessons learned from current demonstrations.

In almost all demonstrations, benefits counseling and case management were offered in conjunction with job placement and VR services. Examples of VR services included unpaid career and work exploration, job training, service learning, job shadowing, work sampling, and job interview training (Honeycutt et al. 2018) and soft skills training such as communication skills, time management skills, and networking skills (DOL 2018). There is strong evidence to suggest that the combination of benefits counseling and VR services results in better employment outcomes (employment status, earnings, hours worked) when compared to benefits counseling or case management in isolation. Future demonstrations should continue offering these two services in tandem.

The timing and nature of benefits counseling is of utmost importance. Preliminary evidence suggests that beneficiaries who waited too long (from the time of application to VR services) to receive benefits counseling and employment services tended to earn less and work fewer hours overall (Honeycutt and Stapleton 2013; Martin and Sevak 2020). Chapter 5 in this volume provides detailed evidence on the importance of early

[17] See Tremblay et al. (2004) for additional details on how the samples for the comparison groups were defined and constructed.

intervention programs that target individuals who are at risk and not yet detached from employment. Of the types of benefits counseling, a tailored, coaching-based intervention was generally found to be more effective than an information-sharing intervention. For example, in a coaching-based approach, beneficiaries are guided through various strategies to track their income, such as the use of calendar tools to track Extended Period of Eligibility (Chambless et al. 2011). Where benefits counselors provided tailored counseling and helped participants develop employment-related goals and actionable steps (e.g., PROMISE and AB Plus), increases in earnings and employment were more likely.

In comparison to other demonstrations, BOND arguably had the most extensive benefits counseling in its provision of WIC and EWIC services. An important finding of BOND was that WIC and EWIC counselors reported feeling burdened by providing post-entitlement services to beneficiaries. Another issue with the implementation of BOND was inadequate training and lack of awareness among counselors of how BOND offset worked. This resulted in a certain level of confusion among beneficiaries about qualifying for and participating in BOND. Improved training and continuing education opportunities could help counselors be better prepared to deliver new programs.

There was wide variation in how benefits counseling was defined and how programs were structured within SSA demonstrations. The duration and intensity of benefits counseling varied considerably or could not be clearly documented between sites. In general, there was a lack of adequate information regarding what happens in a counseling or case management interaction. Even in programs that allowed prolonged engagement and had documentation of the hours spent in counseling and case management, there was scant publicly available information (except BOND) on the content of the sessions.

Last, the availability of integrated social and health care services provides a more optimal environment for implementing benefits counseling programs. There are several examples of integrated care models that support social and health care needs of adults with disabilities and older adults. For example, empirical research on supported employment services integrated with psychiatric care has shown them to increase employment. The integrated ACT model of case management, where services are provided in teams of medical and social service professionals, is highly successful (Bond et al. 2001; Burns et al. 2001; Dixon 2000). For individuals with a dual diagnosis of mental health and substance use issues, peer-led, community-based, integrated programs are considered a best practice (Coldwell and Bender 2007; Bond et al. 2001; Dixon 2000).

FUTURE RESEARCH & LEARNING AGENDA FOR SSA

SSA administers safety net programs that provide income security to families and individuals based on age, disability status, or work credits. These programs operate in a constantly changing environment of economic trends, labor markets, demographic

shifts, and government priorities (Autor, Maestas, and Woodberry 2020). SSA's demonstrations offer rich data and contextual information to understand how, when, and what works in benefits counseling and case management for beneficiaries. This section uses evidence from past and current SSA demonstrations and other empirical research to inform SSA's future policy agenda.

Defining and Operationalizing Benefits Counseling and Case Management through Fidelity Metrics

Establishing benchmark parameters for the content and structure of benefits counseling and case management could be a critical step that enables more accurate monitoring of SSA demonstrations. One way of achieving this is through development and implementation of treatment fidelity metrics for benefits counseling and case management. Treatment fidelity was a critical component of MHTS and the newer SED. Fidelity of intervention is a critical component of determining intervention effectiveness; it is a systematic approach to evaluate and document adherence to the intervention as it was intended. In other words, fidelity is the extent to which an intervention, when implemented, is true to the underlying therapeutic principles (Teague, Bond, and Drake 1998; Waltz et al. 1993). Treatment fidelity was a critical component of MHTS and the newer SED.

Fidelity assessments allow researchers and practitioners to engage in reflective appraisals of the intervention. Fidelity assessments can also inform replicability of findings (or lack thereof) across repeated research and implementation efforts and to isolate intervention program components, as in differentiating between case management and information and referral. For example, counseling theory suggests that active ingredients for any counseling program should include rapport and trust building and deep engagement between the counselor and the beneficiary or recipient. In the absence of opportunities to engage deeply and problem solve collaboratively, counseling sessions are reduced to information and referral sessions. The impact of building trust and rapport was further demonstrated by the success of coordinated, warm handoffs over written benefits summaries within PROMISE (Schlegelmilch et al. 2019).

Developing metrics for and documenting quality indicators for benefits counseling and case management beyond the number of sessions or frequency of contact can provide additional insights into the effectiveness of those services. Such benchmarks for fidelity should specify minimum criteria for content and structure. Typically, fidelity measures include two sets of criteria: (1) structure and process of intervention delivery (the *how*) and (2) content integrity and differentiation of intervention components (the *what*) (Feely et al. 2018). The criteria related to structure and process address the context in which the intervention happens. For example, the number of counseling sessions ("dosage") and whether counseling is provided online or face-to-face ("mode") are structural aspects of intervention fidelity, whereas the specific information or knowledge shared are the active ingredients or core content of

the intervention. Incorporating fidelity measurements into a process and impact evaluation will help SSA evaluate the quality of benefits counseling and case management.

Study Design–Related Issues

Although benefits counseling and case management were critical components of several SSA demonstrations, the unique impact of the programs remains unknown. Because the evaluations were not designed with the specific goal of isolating the effectiveness of case management or benefits counseling (except AB and BOND), the overall effectiveness of these two services in isolation remains unclear. Future demonstrations should consider multi-arm studies or factorial designs of small pilot populations that offer benefits counseling or case management (in conjunction with VR services) tested against other approaches such as benefits offset, work incentives, and the like.

Multi-arm and factorial designs with small clusters of matched or randomly selected beneficiaries or recipients could be helpful in differentiating between small variations in benefits counseling or case management and solidifying the isolated or unique effectiveness of either. For example, multi-arm and factorial designs can be used to simultaneously compare a four-week versus six-week benefits counseling intervention or a telephone versus face-to-face intervention against a single control group. A second approach would be to explore testing multiple clusters of subpopulations sequentially to allow implementing and testing incremental changes to case management and benefits counseling. For example, case management or benefits counseling services can be tested among multiple subgroups based on type of impairment (physical versus sensory versus cognitive) or level of motivation and job readiness.

The evaluation designs in SSA demonstrations included both random sampling and volunteer participation. A recruitment strategy has implications for demonstrating the overall effectiveness and generalizability of findings. Random sampling offers protection against biases in the characteristics of participants in a study. Participants who self-select or volunteer for demonstrations might be highly work oriented or more motivated to return to work. Individuals who turned down enrollment in SED, for example, cited general lack of interest, assumed they cannot work, and cited health issues and other life obligations more frequently than did individuals in the treatment group. Most demonstration evaluations used non-representative selection to recruit participants.

Future programs should consider the effect of non-representative sample selection on employment outcomes and adjust for the same using study design features or statistical controls.

A second issue in the design of demonstrations is the lack of clarity in explaining the causal mechanism between benefits counseling and/or case management and employment outcomes such as earnings or hours worked. A causal mechanism is a

postulated set or sequence of events that links a particular event to an outcome. Causal mechanisms are helpful in explaining *why* certain things happen and to uncover the underlying processes that cause the change (Imai et al. 2013). In social and behavioral science research, a weak causal link between the hypothesized intervention and proposed outcomes can undermine the external validity of a study. If the primary purpose of benefits counseling and case management is to provide accurate information and to monitor use of benefits, the proximal or direct outcome of such intervention is likely to be an increase in knowledge or awareness of benefits and work incentives. Increased knowledge and awareness of (loss of) benefits may motivate some individuals to seek or sustain employment or increase their work hours and earnings, but such outcomes should be considered an indirect effect rather than a direct result of benefits counseling.

Future SSA programs should choose appropriate outcomes for evaluation by carefully considering the proximal outcomes of benefits counseling and case management. This can be accomplished by using theory-based evaluation methods where each component of the benefits counseling and case management intervention is mapped to potential outcomes and tested statistically or by using a case-based approach.

Motivational Interventions

Fear of losing benefits and negative belief systems continue to be a barrier to employment. The prevalent belief of beneficiaries and recipients that being eligible for SSDI or SSI I payments means they are ineligible to work is a persistent barrier to seeking employment. Similarly, motivation to return to work is a strong predictor of return to work. Benefits counselors could be engaged to provide motivational interventions that address beneficiaries' and recipients' negative thought processes and belief systems. For example, benefits counseling could be combined with motivational interviewing or cognitive behavioral therapy techniques to address negative belief systems about inability to work. Similarly, including a plan for actionable change using principles of behavioral economics could be tested. The use of actionable goals is supported by some preliminary evidence from PROMISE and AB Plus. SSA has recently announced that it will conduct an Exits from Disability study, which plans to incorporate motivational interviewing for a sample of SSDI beneficiaries and SSI recipients who exit SSA disability programs because they experience medical improvement.

Acknowledging and Evaluating Work Behaviors Based on Career Trajectories of SSDI Beneficiaries and SSI Recipients

Evidence from VR employment counseling suggests that the pathway to economic self-sufficiency is not linear, especially for individuals with severe limitations. The journey to economic self-sufficiency occurs in intermittent phases,

through participation in apprenticeship, temporary or seasonal work, part-time work, shadowing, temporary staffing, gig work, and so on (Kosciulek 2004). SSDI beneficiaries and SSI recipients who are re-entering the workforce might need several years to re-establish and stabilize themselves in a job and seek higher earnings through increased hours or career advancement. Short-term follow-up studies within the timeframe of 6-18 months might not capture these longer-term outcomes.

Youth with disabilities, who are transitioning to employment, could do so by engaging in internships, apprenticeships, and temporary jobs. Policymakers might consider embedding benefits counseling within programs that target internships or apprenticeships as an early intervention approach for youth in transition. Because apprenticeships and internships are an important milestone experience for youth with disabilities, embedding benefits counseling within them could build awareness early on and set youth on a trajectory for long-term economic self-sufficiency (Iwanaga et al. 2021). There is some evidence supporting the effect of benefits counseling on transition-age youth; future demonstration efforts could be focused on implementation or scaling-up of such services rather than on additional effectiveness or impact evaluation.

The nature of the jobs undertaken by SSDI beneficiaries and SSI recipients should be considered as a potential confounding variable. Beneficiaries and recipients who are employed in jobs that offer a natural pathway or trajectory up the career ladder could have greater potential for increasing earnings through career self-management and advancement. For example, beneficiaries and recipients who work in small businesses with limited staffing needs and positions might not have opportunities to advance in the short term. A vast majority of individuals with disabilities do not have any opportunity to engage in mentoring and career planning (Kulkarni and Gopakumar 2004).

Benefits counseling and case management could be supplemented by career planning and coaching services once a beneficiary or recipient is successfully placed in a job. Career planning and advancement is a process of adjustment an individual goes through to achieve satisfactory job performance and growth (King 2004; Kossek et al. 1998; Kulkarni and Gopakumar 2014). Sustaining and advancing in a job requires active planning and participation in the form of developing new job skills, networking, seeking feedback and advice, and developing insights into one's own career performance and aspirations (Claes and Ruiz-Quintanilla 1998; Kulkarni and Gopakumar 2014; Seibert, Kraimer, and Crant 2001). Sustaining or advancing in a job requires a different set of skills than does getting hired or placed in a job and a different type of case management and follow-up.

SSI recipients and SSDI beneficiaries might benefit from an extended model of support that does not end with benefits management or job placement but rather extends to career coaching for advancement and growth. For example, sustaining in a job requires demonstrating consistent work ethics, social interaction skills, and adequate time and task management. Advancing at work requires demonstrating

initiative, handling additional job task responsibilities, and self-advocating. Beyond job placement services, SSA might consider mentorship or coaching programs that support development of these behaviors at work. Future SSA demonstrations could consider conducting outreach to employers, human resource management professionals, and business leaders to facilitate work behavior outcomes that are consistent with developmental patterns in career trajectories of workers with disabilities.

Duration of Follow-Up

Retaining employment and increasing earnings potential for employed beneficiaries and recipients could take several years. The short duration of demonstrations makes it challenging to observe any long-term or distal outcomes such as those. Tremblay et al. (2004, 2006) used a four-year time frame following benefits eligibility determination to demonstrate improvements in employment status and wages earned. A longer follow-up duration might allow sufficient time for some of these career development activities to transpire. Such long-term follow-up activities can extend beyond the life of the demonstration itself.

Mediating Role of Work Incentives

SSI recipients and SSDI beneficiaries rely on a wide range of supports to sustain and advance in their jobs. The use of federally funded work incentives such as impairment-related work expenses or Plan to Achieve Self-Support can vary over the course of a beneficiary's or recipient's work life. Further longitudinal investigation of the timing and intensity of such services as mediators of the use of workplace accommodations and advancement could reveal new trends in how beneficiaries achieve self-sufficiency (Iwanaga et al. 2021).

Financial Literacy Training

Reaching economic self-sufficiency requires both income generation and asset building. Current SSA programs focus on income generation through finding and maintaining employment. Asset building by saving for emergencies and unforeseen circumstances can be considered a complementary strategy for reaching economic self-sufficiency. SSI recipients and SSDI beneficiaries have access to ABLE accounts to save for expenses related to living with a disability such as purchase of assistive technology, payments for housing, accessible transportation, and the like. In general, savings in ABLE accounts do not affect eligibility for SSI, Medicaid, SNAP, and the like.

Financial literacy training combined with benefits counseling can provide beneficiaries, recipients, and their families with tools to demystify the larger picture of economic self-sufficiency. Families and individuals with low incomes may lack the necessary financial literacy skills required to make informed financial choices.

Coaching financial literacy skills such as budgeting and money management can help beneficiaries and recipients feel more secure about their economic well-being and, in the long term, build assets. Access to financial literacy training that will propel individuals to save money in the long term is another way beneficiaries and recipients can build assets and become more economically secure. Though financial literacy training is included in some SSA demonstrations, an increased emphasis on the same and rigorous evaluation of the impact of such training will add to the existing knowledge base on this topic. Financial literacy training has its limitations, however; it may not be relevant to families who do not have the financial means to save.

Targeted Case Management Services

There is moderate to strong evidence from empirical research including SSA demonstrations that case management is an effective practice when implemented with high fidelity. Case management is especially effective in improving employment outcomes for individuals with mental health conditions. Case management offered within the context of high-fidelity supported employment programs also has been demonstrated to be effective for them. Data from Project NetWork highlights the cost-prohibitive nature of such services. Future implementation of case management targeted to a section of the beneficiary population who are at high risk for not returning to work or are least job ready could be a fiscally responsible approach.

Embedding Services in Integrated Health Systems

In the United States, health care and long-term services/supports have historically been delivered through separate and siloed channels. Health care organizations provide medical care whereas community-based organizations provide services that address social determinants of health factors (transportation, caregiver supports, Meals on Wheels, etc.). The Administration for Community Living (2020) has recently engaged in strategic planning to integrate health care and social services for individuals with disabilities and for older adults. Future SSA demonstrations could consider embedding employment support services within integrated programs to address health care needs and social determinants of health. For example, ACT programs provide integrated health and social services to individuals with mental health issues. Commonly referred to as the *hospital without walls* approach (Dixon 2000; Ellenhorn 2005), ACT teams deliver health and social care in integrated teams in the community rather than in residential hospital settings. A similar strategy could be used to embed benefits counseling and case management within integrated health and social service teams.

CONCLUSIONS

Many Americans with disabilities are striving to work and overcoming barriers to reach economic self-sufficiency. Benefits counseling and case management have been

characterized as essential services to assist them to return to or seek employment. However, evidence supporting the impact of these strategies on improving work outcomes and earnings for SSI recipients and SSDI beneficiaries is, at best, weak to moderate. This could be a function of the heterogeneous nature of the population targeted, variations in the content and structure of benefits counseling and case management programs, or duration and intensity of services provided.

Based on the available evidence, it is challenging to disentangle the unique impact of benefits counseling or case management from other services that were provided as part of SSA demonstrations. Outcome and impact evaluations reported by most SSA demonstrations consolidate and comingle multiple services. Future SSA demonstrations should clarify the scope, intensity, and frequency of benefits counseling and case management and examine their unique impacts through long-term follow-up studies.

Chapter 8

Comment

John Kregel
VCU National Training and Data Center
Virginia Commonwealth University

Dr. Sundar (in "Benefits Counseling and Case Management") provides an excellent analysis of the use of benefits counseling as a component of Social Security Administration (SSA) demonstrations. She documents the differences between benefits counseling in early 2000s demonstrations, such as the State Partnership Initiative (SPI), which occurred as SSA was launching the Benefits Planning, Assistance, and Outreach (BPAO) program, and later demonstrations, such as the Benefit Offset National Demonstration (BOND) and Promoting Opportunity Demonstration (POD), which were developed to be comparable to the current Work Incentives Planning and Assistance (WIPA) model. The WIPA model was redesigned by SSA in 2006 in response to shortcomings in the BPAO program and prioritizes the delivery of benefits counseling to SSDI beneficiaries and SSI recipients who are employed or have a job offer pending (Kregel and O'Mara 2011).

Including benefits counseling services in demonstrations in a manner similar to those in the WIPA program provides support to the treatment group participants throughout the intervention and enables SSA to assess the feasibility of widespread implementation of policy changes or program waivers. The comments below describe:

- the importance of high-quality benefits counseling services in SSA demonstrations; and
- strategies that should be used to standardize benefits counseling interventions in multi-site demonstrations.

IMPORTANCE OF BENEFITS COUNSELING IN SSA DEMONSTRATIONS

As described in Chapter 8, benefits counseling services are often combined with other interventions in SSA demonstrations. In some demonstrations, such as SPI and Promoting Readiness of Minors in Supplemental Security Income (PROMISE), the design allowed variation across sites in the job descriptions of benefits counselors and the manner in which benefits counseling was combined with other interventions. This approach enabled individual sites to develop interventions that responded to the needs of participants and the unique characteristics of the state/local service delivery system, but limited the extent to which results could be combined across sites.

In contrast, in the BOND and POD projects, SSA designed the benefits counseling intervention to be comparable to the WIPA program. The design established specific performance measures that required all benefits counseling services to meet basic quality standards. Though a proposed policy change, such as the gradual benefit offset

in BOND and POD, may seem simple and straightforward, SSA conducts demonstrations in the context of a highly complex regulatory system. Benefits counselors must be able to assist beneficiaries to navigate the SSDI program rules addressing the effect of increased earnings on benefit amounts and program eligibility, use of work incentives, relationship between work and continued health care coverage, availability of other federal and state-specific benefits, and the unique situations of concurrent beneficiaries and self-employed individuals.

In summary, if benefits counselors deliver inaccurate or incomplete information to demonstration participants, it can have negative consequences for SSDI beneficiaries and SSI recipients. Effective benefits counseling requires work incentives counselors to possess a combination of detailed technical knowledge, high-level counseling skills, and ability to accurately describe complex information to beneficiaries and recipients in a way that will enable them to make confident decisions about their careers and health insurance coverage. The development of rigorous performance standards for the delivery of benefits counseling services should continue to be the standard for future SSA demonstrations that test new policies or programs.

STANDARDIZING BENEFITS COUNSELING INTERVENTIONS IN MULTI-SITE DEMONSTRATIONS

Designing and implementing effective benefits planning components of SSA multi-site demonstrations require the development of replicable service protocols, rigorous training for work incentives counselors, and continuous technical assistance to maintain service integrity. A lack of standardization can make it difficult to aggregate data across sites or assess the use of evidence-based or promising practices. For example, in the context of POD, standardization efforts focused on development and monitoring of service delivery protocols; rigorous training of work incentives counselors; and ongoing, intensive technical assistance.

Standardization of Service Delivery Protocols

The POD benefits counseling intervention was based on the development of SSA-approved service delivery protocols that cover the following service components: onboarding and engagement, earnings and benefits verification, counseling on the specific alternative rules of the demonstration, Benefits Summary and Analysis report preparation, referral for employment services and supports, and off-boarding (return to standard SSDI rules, if included as a part of the demonstration). In POD, ongoing monitoring of the implementation of these protocols made it possible to assess the effectiveness of the intervention across multiple sites.

Standardization of Work Incentives Counselor Training

All POD counselors not previously certified as work incentives counselors completed a formal, competency-based training program, based on rigorous

assessments, prior to beginning services. In addition, all counselors were required to complete a comprehensive training module addressing the rules and procedures specific to the demonstration.

Standardization of Technical Assistance

SSA required the POD implementation contractor to provide ongoing technical assistance to each individual site manager and individual work incentives counselors. Technical assistance included monthly webinars with site managers and work incentives counselors, semi-annual site visits to each site designed to enhance compliance with all service delivery protocols, and monthly calls with individual work incentives counselors to conduct case reviews on individual participants.

CONCLUSION

As documented in Chapter 8, variation in the delivery of treatment group interventions sometimes makes it difficult to aggregate data across multiple sites in SSA demonstrations. In designing the POD project, SSA sought to standardize treatment interventions across program sites by developing detailed service delivery protocols, providing rigorous training for demonstration staff, and monitoring site performance throughout all phases of the intervention. SSA's continued use of these standardization strategies can increase the overall fidelity of the interventions and promote the use of promising or evidence-based service practices in SSA demonstrations.

Chapter 8

Comment

Leslynn R. Angel
Michigan United Cerebral Palsy

Over the past 40 years, the Social Security Administration (SSA) has conducted many demonstration projects that have incorporated benefits counseling and case management. For most of those demonstrations, the goal has been to identify ways to reduce reliance on benefits, decreasing participation in Supplemental Security Income (SSI) and Social Security Disability Insurance (SSDI). Although there has been a recent decline in participation, there continues to be a lack of meaningful changes in unemployment for people with disabilities. According to data from the US Bureau of Labor Statistics, in 2020 the unemployment rate for people with disabilities was at 12.6 percent, the highest in seven years.

The Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act) was signed into law to increase the options for individuals with disabilities who wished to return to work. Through the Ticket to Work program, benefits counseling was recognized as a core service for those receiving SSI and SSDI benefits. Benefits counseling has transitioned from the Benefits Planning Assistance, and Outreach (BPAO) program, which ensured beneficiaries and recipients were receiving accurate information, to the current Work Incentives Planning and Assistance (WIPA) program. WIPA focuses on providing benefits counseling to those who are working or have an active work goal.

One of the challenges with the WIPA program is funding. Programs in 2021 continue to be funded at the same level of the initial BPAO projects in 1999. The WIPA program was tasked by SSA to "disseminate accurate information to beneficiaries with disabilities…about work incentives programs and issues related to such programs." The ultimate goal of the assistance was to "assist SSA beneficiaries with disabilities succeed in their return-to-work efforts" (SSA 2006).

Over the years it has been difficult for WIPA to address *all* employment barriers faced by SSDI beneficiaries and SSI recipients with disabilities, such as work disincentives contained within SSA, overpayments and other benefit programs, employer reluctance to hire them, fears of losing health care, or lack of service providers to assist them in acquiring the skills they need to find and retain employment. What WIPA is able to address are the barriers to work due to beneficiaries' lack of understanding of work incentives or inability to connect with resources to support their employment.

The implications that Ticket to Work had for many SSI recipients and SSDI beneficiaries were major. The program has given them the opportunity to have greater control and choice of their path to work.

SSI and SSDI provide economic security for many who are living below the federal poverty level. After waiting many months, sometimes years, before being accepted to receive benefits, the idea of working is a scary reality for most. Many people are afraid to go to work or have the mindset that they cannot work that promoting employment and economic stability from the beginning of their participation in SSDI or SSI is a challenge.

Benefits counseling and case management have been critical components of SSA's demonstrations. From Dr. Sundar's discussion of the Benefit Offset National Demonstration, Promoting Readiness of Minors in SSI, State Partnership Initiative, Accelerated Benefits, Mental Health Treatment Study and Project NetWork, we discover that supporting various populations requires different approaches. Building trust and a working relationship are also critical.

Therefore, using a one-size-fits-all approach to supporting individuals receiving benefits will more than likely not gain positive results. We learned that incorporating a person-centered approach based on a person's individual circumstances will likely garner the best results. A person-centered approach is where the person is placed at the center of the service; the focus is on the person and what they can do, not on their condition or disability. Support should focus on achieving the person's aspirations and be tailored to their needs and unique circumstance.

There is ample evidence that incorporating case management works well for youth and those with mental health–related disabilities. We also know that navigation of the complex Social Security rules is very difficult for most. Trained benefits counselors are critical to provide much needed information. As Dr. Sundar discussed, some individuals, such as those with mental illness, are aided most by benefits counseling and case management paired with Vocational Rehabilitation (VR) services. She also concluded that case management might not be relevant for all beneficiaries and recipients.

If we place a greater emphasis on subpopulations and identifying what works and does not work, we will likely have a greater impact on service delivery. Dr. Sundar also indicated that financial literacy is another tool to promote self-sufficiency and that could instill the desire to work or return to work.

Timing is important in relationship to benefits counseling and VR services. For example, working with transition-age youth and incorporating benefits counseling as part of the transition plan will plant an early seed for youth who will soon be exiting the educational system, making the handoff to VR a more natural progression to independence. Similarly, encouraging SSDI beneficiaries and SSI recipients at the onset of receiving benefits that employment is an option and making benefits counseling available immediately can support their transition back to work.

It is promising that SSA continues to implement demonstrations that focus on incorporating case management and benefits counseling, but we need to take a deeper dive at the underlying issues surrounding unemployment and the disincentives of returning to work.

Chapter 9

Lessons from Implementation

Michelle Wood
Abt Associates
Debra Goetz Engler
Social Security Administration[1]

The demonstrations conducted by the Social Security Administration (SSA) typically use rigorous impact evaluations to estimate intervention effects and process analyses[2] to examine how interventions are implemented. This chapter considers the second component. The demonstrations' process analyses have focused on (1) describing the implementation (both how the interventions operate and service receipt); (2) assessing fidelity to the intervention model and reasons for deviations; (3) documenting contextual factors such as the labor market, economic conditions, and social service systems; (4) describing the counterfactual condition; and (5) identifying lessons learned and promising practices.

Understanding implementation is vital to interpreting effects; impact analyses alone cannot explain why an intervention does or does not achieve intended results. For example, a demonstration might not produce effects because the intended intervention was poorly implemented (Epstein and Klerman 2012). It is also possible that favorable effects might be found when the intervention implemented differs in important ways from original plans. Research relying on naturally occurring variation in implementation conditions shows that implementation affects outcomes (Bloom, Hill, and Riccio 2003; Durlak and DuPre 2008).

The demonstrations' evaluations are most useful to policymakers if they document how implementation occurred and assess whether the intervention as implemented represents a reliable test of the intended model. If evaluation findings show that an intervention is effective, implementation findings can shed light on how to best replicate and scale it. Even when evaluations do not find favorable effects or interventions are poorly implemented, a careful process analysis can help to identify strategies for strengthening implementation in a future replication or bolstering the program design to increase its effectiveness.

The chapter's first section addresses the implementation mechanics of recruitment and enrollment. Findings about recruitment shed light on likely take-up of different kinds of services and financial incentives and can guide future efforts to conduct

---

[1] The views expressed in this chapter are those of the authors and do not necessarily represent the views of the Social Security Administration or the US federal government.

[2] Throughout we use the term *process analysis* to refer to qualitative research that examines the implementation of interventions. A commonly used synonym in the field of program evaluation for this type of research is *implementation analysis*.

outreach and encourage service participation. The second section examines service delivery. It highlights lessons about how services have been delivered, the extent to which participants use the services offered to them, factors that lead to variation in implementation across program locations, and implications for interpreting impacts. The chapter's third section summarizes the overarching lessons and directions for the future.

LESSONS ABOUT RECRUITING AND ENROLLING DEMONSTRATION PARTICIPANTS

To draw lessons about recruitment and enrollment we focus on the 12 demonstrations shown in Exhibit 9.1. The demonstrations vary in whom they tried to reach and what type of assistance they offered. However, all shared the goal of recruiting an appropriate sample sufficient in size to detect meaningful effects or to support other analyses that can inform policy. We begin by examining the results of recruitment. We compare the response to outreach and recruitment efforts across the demonstrations and the methods the demonstrations have used to recruit participants. We then explore findings from the demonstrations about how volunteers compare to non-volunteers.

Exhibit 9.1. Demonstrations Reviewed to Identify Lessons about Recruitment/Enrollment

	Target Population and Enrollment Period	Intervention	Data Source(s) for Identifying Population
Early Interventions—Before SSDI/SSI Application			
Demonstration to Maintain Independence and Employment (DMIE)[a]	Working-age adults with chronic conditions who are not SSI/SSDI applicants or recipients/beneficiaries 2006–2008	Health insurance and employment services	Varies by program: employers, state insurance programs, public health systems
Retaining Employment and Talent After Injury/Illness Network (RETAIN)	Working-age adults with illness or injury who are not SSI/SSDI applicants or recipients/beneficiaries Phase 1: 2019–2020 Phase 2: 2021–2022	Return-to-work coordination; occupational health best practices; workplace-based interventions; training and rehabilitation services	Varies by program: employers, state insurance programs, workers' compensation; health care systems
Supported Employment Demonstration (SED)	Denied SSDI and concurrent applicants who allege mental health condition, with an interest in working, except for those: (1) incarcerated; (2) with cognitive impairments; (3) enrolled in an employment and training program 2018–2019	IPS model of supported employment with integrated medical and behavioral health care and financial assistance for care not covered by individuals' health insurance plans	SSA administrative data; additional eligibility screening

	Target Population and Enrollment Period	Intervention	Data Source(s) for Identifying Population
Broad Appeals to SSDI Beneficiaries and SSI Recipients			
Benefit Offset National Demonstration (BOND)	Stage 1: SSDI-only and concurrent Stage 2: SSDI-only 2011–2012	$1 for $2 benefit offset; benefits counseling	SSA administrative data
Project NetWork	SSDI-only and concurrent; SSI recipients; SSI applicants 1992–1993	Case management	SSA administrative data (SSDI beneficiaries and SSI recipients); SSA claims representatives (SSI applicants)
Promoting Opportunity Demonstration (POD)	SSDI-only and concurrent 2018	$1 for $2 benefit offset; benefits counseling	SSA administrative data
Specialized Services for Specific Groups of SSDI Beneficiaries and SSI Recipients			
Accelerated Benefits (AB)	New SSDI beneficiaries in the Medicare waiting period who do not have health insurance 2007–2009	Health insurance or health insurance plus progressive goal attainment; benefits counseling; medical case management	SSA administrative data; additional eligibility screening
Mental Health Treatment Study (MHTS)	SSDI beneficiaries with schizophrenia or affective disorders except for those: (1) in nursing homes; (2) with legal guardian; (3) with life-threatening or terminal illness; (4) receiving supported employment in past 6 months; (5) with a competitive job 30 days before enrollment 2006–2007	IPS model of supported employment; systematic medication management; nurse care coordination	SSA administrative data; additional eligibility screening
Promoting Readiness of Minors in Supplemental Security Income (PROMISE)	Youth receiving SSI, ages 14–16 2014–2016	Case management; benefits counseling and financial literacy training; career and work-based learning for youth and family members; parent training or information	SSA administrative data

	Target Population and Enrollment Period	Intervention	Data Source(s) for Identifying Population
State Partnership Initiative (SPI) New York WORKS project	SSI recipients with psychiatric diagnosis over age 21 2000–2003	Benefits counseling; employment coordination	SSA administrative data
Transitional Employment Training Demonstration (TETD)	SSI recipients with intellectual disability, ages 18–40 1985–1986	Placement in competitive jobs; on-the-job training; postemployment and job retention services	SSA administrative data; additional eligibility screening
Youth Transition Demonstration (YTD)	Youth receiving SSI, ages 14–25 2006–2008	Case management; benefits counseling and financial literacy training for youth and parents; career and work-based learning	SSA administrative data

Key: IPS=Individual Placement and Support. SSDI= Social Security Disability Insurance. SSI= Supplemental Security Income.

Source: Authors' summary of demonstration reports. AB: Michalopoulos et al. (2011). BOND: Gubits et al. (2018a/b). DMIE: Gimm et al. (2009); Whalen et al. (2012). MHTS: Frey et al. (2011). POD: Hock et al. (2020). Project NetWork: Kornfeld and Rupp (2000). PROMISE: Anderson et al. (2018); Honeycutt, Gionfriddo, Kauff, et al. (2018); Kauff et al. (2018); Mamun et al. (2019); Matulewicz, Katz, et al. (2018); McCutcheon et al. (2018); Selekman et al. (2018). SED: Taylor et al. (2020). SPI New York WORKS: Ruiz-Quintanilla et al. (2006). TETD: Thornton and Decker (1989). YTD: Fraker, Mamun, et al. (2014).

[a] Centers for Medicare and Medicaid Services sponsored this demonstration.

Findings about recruitment and enrollment provide direct feedback on the response to the recruitment methods used and demand for the assistance being offered. The findings can also tell us about the potential interest in different interventions among various types of individuals.[3] Findings about which types of outreach strategies did and did not work can hold lessons for designing future demonstrations. Findings about recruitment might also hold lessons for operating SSA's ongoing Ticket to Work (TTW) and Work Incentives Planning and Assistance (WIPA) programs, which also have goals to promote attachment to the labor force. Findings about recruitment and enrollment challenges might suggest ways to focus future outreach efforts or to increase participation among groups with the greatest policy interest.

[3] Recruitment activities end when demonstrations achieve the target sample size; therefore, the enrollment rates might indicate a lower-bound estimate of the level of interest in the interventions. In several demonstrations, enrollment was also time limited. Given this, the proportion of eligible individuals who enroll might not provide evidence about the maximum potential interest in the intervention.

Recruitment Results

This section compares the results of outreach and recruitment. The comparison yields three principal lessons about (1) the response to outreach and recruitment, (2) the use of various recruitment methods, and (3) special considerations for early interventions.

SSA demonstrations have successfully recruited both broad and specific target populations; but in most cases, those targeting narrowly defined groups have achieved the strongest response to outreach.

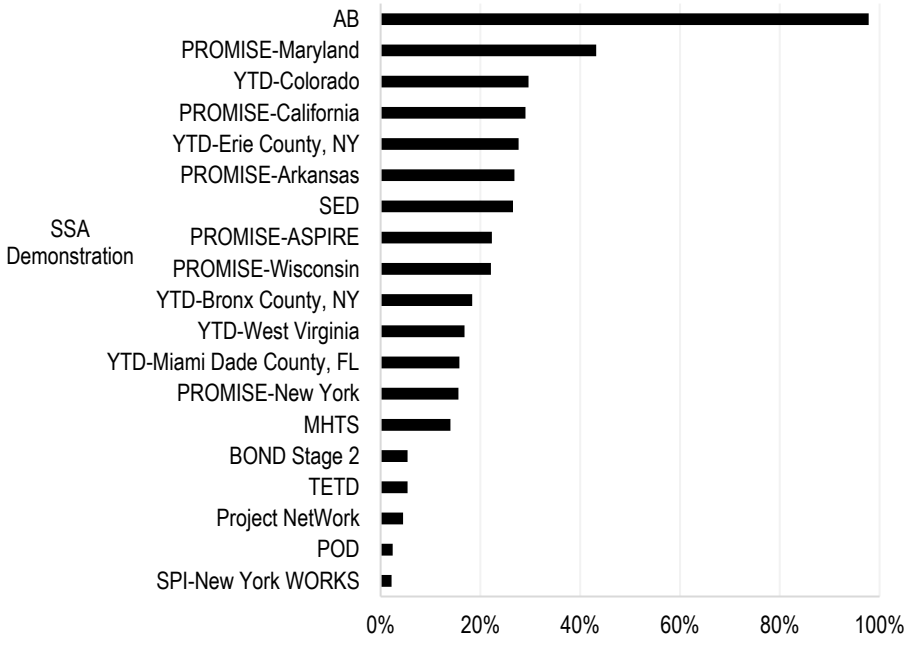
Overall, new Social Security Disability Insurance (SSDI) beneficiaries, denied SSDI and concurrent applicants, and youth have been more likely to volunteer than existing SSDI beneficiaries and Supplemental Security Income (SSI) recipients. The stronger response may reflect the appeal of the intervention (e.g., health insurance or employment services for denied applicants), as broad appeals offering financial incentives (BOND, POD, Project NetWork) attracted the lowest rates of volunteers, between 2.4 and 5.4 percent of disability beneficiaries (Gubits et al. 2018a/b; Hock et al. 2020; Kornfeld and Rupp 2000). As reported by Thornton and Decker (1989) and Ruiz-Quintanilla et al. (2006), two demonstrations targeting narrowly defined groups of SSI recipients, TETD and the New York WORKS SPI project, also had comparatively low enrollment rates (5.4 percent and 2.2 percent). Exhibit 9.2 below displays the recruitment results, expressed as the proportion enrolled of all eligible individuals recruited. (The summary exhibit at the end of this chapter provides additional detail about recruitment results.)

TETD and Project NetWork. SSA's early experiences with national demonstrations showed that it is feasible to conduct outreach with large numbers of SSDI beneficiaries and SSI recipients and secure their participation. TETD offered job placement, on-the-job training, job retention, and waivers of SSI rules.[4] SSA sent invitation letters to 13,800 SSI recipients with intellectual disability (Decker and Thornton 1995; Thornton and Decker 1989). Demonstration intake staff in the eight programs that implemented TETD conducted recruitment. These staff made follow-up phone calls, sent reminder letters to potential volunteers, and engaged with community organizations to inform them about the demonstration. Altogether, 2,404

[4] Thornton, Dunstan, and Schore (1988) describe the waivers that SSA obtained for the TETD project. Three of the waivers allowed demonstration participants to maintain eligibility for SSI benefits while receiving training and working. The first did not count earnings during the demonstration as an indicator of Substantial Gainful Activity (SGA). The second excluded time working from calculations of the Trial Work Period (TWP). The third guaranteed participants a 15-month Extended Period of Eligibility. A fourth waiver excluded earnings that a participant saved from asset limitations in the SSI program. The waivers are documented in *Federal Register* 50, No. 85 (May 2, 1985): 18741-18742.

SSI recipients (17 percent) responded to the initial letter and attended an intake session, and 745 of those solicited (5.4 percent) volunteered.

Exhibit 9.2. Recruitment Results, Percentage Enrolled of Those Eligible, by Demonstration



Source: Authors' summary of demonstration reports. AB: Michalopoulos et al. (2011). BOND: Gubits et al. (2018a/b). MHTS: Frey et al. (2011). POD: Hock et al. (2020). Project NetWork: Kornfeld and Rupp (2000). PROMISE: Anderson et al. (2018); Honeycutt, Gionfriddo, Kauff, et al. (2018); Kauff et al. (2018); Mamun et al. (2019); Matulewicz, Katz, et al. (2018); McCutcheon et al. (2018); Selekman et al. (2018). SED: Taylor et al. (2020). SPI New York WORKS: Ruiz-Quintanilla et al. (2006). TETD: Thornton and Decker (1989). YTD: Fraker, Mamun, et al. (2014).
Note: Project NetWork enrolled a total of 8,248 in the evaluation. Of those, 6,527 (4.5 percent) were enrolled through the outreach and recruitment process. The remaining 1,721 were new SSI applicants recruited by SSA claims representatives.

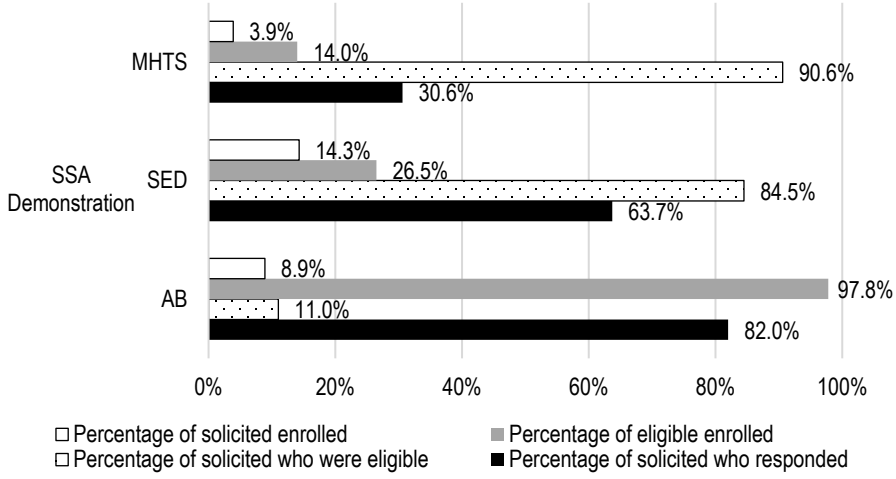
SSA used the experiences in TETD to inform the design of Project NetWork. Project NetWork offered case management and program waivers[5] to a broad population of SSDI beneficiaries, SSI recipients, and SSI applicants regardless of disabling condition. Project NetWork used similar procedures as TETD to conduct

[5] For SSDI beneficiaries, the Project NetWork waivers exempted earnings for a 12-month period when computing TWP months and prevented benefit suspension for those who already had exhausted the TWP. For SSI recipients, the waivers prevented earnings from triggering a medical continuing disability review as would otherwise happen under current-law rules.

outreach to 145,404 SSDI beneficiaries and SSI recipients (Kornfeld and Rupp 2000). SSA sent initial letters to potential volunteers that contained a postcard and instructions that interested individuals return the postcard to complete the enrollment process. Demonstration intake staff conducted in-person information sessions with those who responded, yielding 6,527 enrollees (4.5 percent of those solicited) (Burstein, Roberts, and Wood 1999). Another 1,721 SSI applicants enrolled in Project NetWork in response to outreach conducted by SSA claims representatives, for a total of 8,248 randomly assigned.

Demonstrations Offering Specialized Services to Specific Groups. Newly entitled SSDI beneficiaries and denied SSDI and concurrent applicants have proven easier to contact and more responsive than other groups. Factors influencing this could be that contact information for recent awardees and denied applicants is likely more up to date than for longer-duration beneficiaries, the appeal of the intervention offered, or the recruitment methods used. Drawing on findings reported by Michalopoulos et al. (2011) and Frey et al. (2011), we show details about the recruitment results for the AB and MHTS demonstrations in Exhibit 9.3 below. Exhibit 9.3 also shows results reported by Taylor et al. (2020) for SED, which recruited denied SSDI and concurrent applicants.

Exhibit 9.3. Recruitment Results for Accelerated Benefits (AB), Mental Health Treatment Study (MHTS), and Supported Employment Demonstration (SED)



Source: Authors' summary of demonstration reports. AB: Michalopoulos et al. (2011). MHTS: Frey et al. (2011). SED: Taylor et al. (2020).

In AB, 82 percent of new SSDI beneficiaries sent a mailing about the demonstration completed the initial interview. The high response rate might indicate strong interest in an easily understood service—health insurance. AB also stands out

for its high rate of enrollment. SSA provided administrative data for newly entitled SSDI-only beneficiaries who were entitled for benefits at the initial level and who had at least 18 months remaining in their Medicare waiting period. The recruitment staff mailed letters to these beneficiaries and then made repeated attempts to contact them by phone to complete eligibility determination, informed consent, baseline interview, and random assignment. Once the recruiting staff confirmed their eligibility,[6] 98 percent of those eligible went on to enroll (Michalopoulos et al. 2011), likely reflecting the appeal of the offer of health insurance.

The AB recruitment process also shed light on an open policy question about the level of unmet health insurance needs among new SSDI beneficiaries (Michalopoulos et al. 2011; Weathers et al. 2010). Of the high proportion (82 percent) who responded to initial outreach in AB, only 11 percent of respondents indicated they did not have health insurance.[7]

Two other demonstrations targeted narrowly defined groups of adults. The currently operating SED offers the Individual Placement and Support (IPS) model of supported employment to denied SSDI and concurrent applicants with mental health conditions[8] in two treatment groups. A "Full-Service" group receives IPS with integrated medical and behavioral health care, systematic medication management and nurse care coordination; a "Basic-Service" group receives IPS and other behavioral health services. MHTS offered IPS, along with systematic medication management and nurse care coordination, to SSDI beneficiaries with schizophrenia or affective disorders.[9]

---

[6] To be eligible for the AB demonstration, new SSDI beneficiaries could not be receiving health insurance and could not be institutionalized. Additional screening during recruitment was necessary to confirm eligibility, as SSA administrative data do not provide all the information needed. The recruiting staff asked questions to verify that the beneficiary was not insured, non-institutionalized, and able to answer survey questions. Once staff had confirmed the beneficiary's eligibility, they obtained informed consent, administered the full baseline survey, and conducted random assignment.

[7] The AB demonstration's 12-month survey provides even more insights about access to health insurance for new SSDI beneficiaries. Neither the treatment group nor control group had health insurance at random assignment, but after a year, 40 percent of control group members obtained health insurance. This finding is important context for interpreting impact estimates. It also fills a gap in knowledge about the extent to which SSDI beneficiaries without health insurance obtain coverage.

[8] SED eligibility was limited to denied SSDI and concurrent applicants with mental health conditions who were interested in work and not participating in employment services. SSA administrative data do not verify those eligibility criteria; they were confirmed during the recruitment process.

[9] Eligibility for MHTS depended on a diagnosis of schizophrenia or affective disorder, which can be identified in SSA administrative data; but additional eligibility criteria—absence of specific terminal conditions (AIDS, end-stage renal disease, terminal cancer), no receipt of supported employment in the past six months, and no competitive employment 30 days before enrollment—were verified by talking with the potential enrollee.

As in AB, for SED the research contractor conducted outreach and recruitment. Using local interviewers, the contractor sent letters to denied applicants using lists that SSA generated from administrative data. The recruitment staff then made up to five follow-up calls and up to two home visits to attempt to contact the potential volunteers. After contacting potential volunteers, the recruiting staff screened for additional eligibility criteria (i.e., the denied applicants had to be interested in working and not receiving employment services). Unlike in AB, once determined eligible, potential volunteers for SED had to attend an in-person recruitment information meeting to learn more about the demonstration services and to provide informed consent. In MHTS and SED, the IPS model requires in-person recruitment.

Altogether, 64 percent of the denied SSDI applicants who were contacted responded to the initial outreach conducted in SED (Taylor et al. 2020). Of those who responded and were determined eligible, 27 percent enrolled. Both the initial response and enrollment were higher in SED than in MHTS. In MHTS, 31 percent of the SSDI beneficiaries solicited responded to the initial contact; of those determined eligible, 14 percent enrolled (Frey et al. 2011). The stronger response from denied applicants could reflect that this group recognizes employment support can help them secure paying jobs or seeks to meet critical needs for medical or behavioral health care that treatment group membership provides. In addition, denied applicants may have more recent work experience than do beneficiaries and less time away from the workforce, and that may have motivated their participation. Finally, it is possible that some denied applicants may think that responding to outreach could improve the possibility of a positive decision about entitlement in the future.

Differences in the recruitment processes used in SED and MHTS might also have influenced the stronger response to initial outreach in SED compared to MHTS. As discussed by Frey et al. (2011), in MHTS, designated demonstration operations staff (called "research assistants") conducted outreach and recruitment, using lists of potential volunteers the research contractor developed from SSA administrative data. The research assistants in the sites sent letters and made follow-up calls and visits to contact potential volunteers. Once they contacted a potential enrollee, the research assistant conducted eligibility screening to verify that the beneficiary had not been employed in a competitive job in the past 30 days, did not have a physical health condition that precluded participation,[10] and had not received supported employment in the past six months. Those who passed the eligibility screen then attended two recruitment information group meetings to learn more about the demonstration, provide informed consent, and enroll.

The recruitment results in SED confirmed the feasibility of engaging with individuals before SSDI entitlement and that the level of interest in employment is high among a particular group of denied SSDI and concurrent applicants. However,

[10] Beneficiaries with AIDS, end-stage renal disease, or terminal cancer were excluded from the demonstration.

SED also points to the difficulties in reaching some denied applicants, mostly because of frequent address changes, unreliable phone numbers, and homelessness.

Demonstrations Targeting Youth. To examine recruitment experiences involving youth, we draw on findings from YTD[11] and from PROMISE[12]. YTD and PROMISE offered a variety of case management and career and work-based learning services to youth receiving SSI. Enrollment rates ranged from 16 to 30 percent in YTD and from 16 to 43 percent in PROMISE. These comparatively high enrollment rates for YTD and PROMISE (see Exhibit 9.2 above) are most likely an indicator of the appeal of the specific services (coupled with program waivers in YTD) that were offered and a high level of interest among youth and their families in pursuing employment-enhancing activities.

The response among youth might also reflect interest in employment and perhaps encouragement from families. Responses to the YTD baseline survey, reported by Fraker and colleagues in 2011 and 2012, indicate that enrollees held positive expectations for the future, and these sentiments could have encouraged them to enroll. In all the YTD programs, more than 80 percent of enrollees reported that they expected to work at least part-time in the future. Between 68 and 79 percent in all the YTD sites said they expected to live independently in the future, and between 66 and 97 percent said they expected to continue their education in the future. In YTD, the evaluators concluded that the SSI program waivers—a more generous $1 for $4 benefit offset in the earned income exclusion and an extension of the student earned income exclusion to age 21—encouraged participants to enroll.

YTD and PROMISE used different approaches for recruitment. In YTD, the research contractor conducted outreach and recruitment centrally and by phone using interviewer staff. The contractor sent letters to youth whom SSA identified in administrative data and then followed up with reminder letters and phone calls. Response to the initial contact ranged between 29 and 45 percent across the six YTD programs. Once these recruiters spoke to a potential volunteer, they obtained verbal consent and completed the baseline interview. The recruiters instructed the potential volunteer to sign and return the informed consent form. When the research contractor received the signed consent, it proceeded to conduct random assignment.

Given flexibility to develop local approaches, PROMISE projects used a variety of strategies to enroll the required sample size. Local demonstration staff conducted recruitment in all except the Maryland project, where the state engaged a local contractor to conduct recruitment and enrollment. In PROMISE, most projects sent enrollment packets (with information about the services, informed consent, and

---

[11] Reported by Fraker, Baird, et al. (2011); Fraker, Black, Broadus, et al. (2011); Fraker, Black, Mamun, et al. (2011); Fraker, Baird, et al. (2012); Fraker, Honeycutt, et al. (2012); Fraker, Mamun, et al. (2012); and Fraker, Mamun, et al. (2014).

[12] Reported by Anderson et al. (2018); Honeycutt, Gionfriddo, Kauff, et al. (2018); Kauff, Honeycutt, et al. (2018); Mamun et al. (2019); Matulewicz, Katz, et al. (2018); McCutcheon et al. (2018); and Selekman et al. (2018).

instructions about how to enroll), followed by phone calls, texts, and follow-up letters, to engage potential participants. One project modified its approach to add an initial postcard prior to the first letter to increase brand recognition. Most projects engaged with community stakeholders such as schools, child welfare agencies, and social workers to inform them about the program. Projects also held community events to publicize their programs and increase awareness. The enrollment results in PROMISE also reflect relatively intense recruitment efforts. The average number of contacts the projects reported making with the enrolled group members ranged from 2.5 to 6.2.

Demonstrations with Broadly Defined Target Populations. The BOND[13] and POD projects targeted broad cross sections of the SSDI beneficiary caseload (and in POD, SSDI/SSI concurrent beneficiaries) with offers of alternative SSDI earnings rules intended to encourage work. In BOND, after a letter and up to five follow-up letters, phone contacts, and an in-person enrollment meeting, 5.4 percent of those solicited enrolled in the second (voluntary) stage of the demonstration (Gubits et al. 2013). Because BOND was a test of a national policy, a key objective was to evaluate the results of uniform recruitment procedures applied consistently throughout the BOND sites. The idea was to learn about interest in a benefit offset among a large population of potential participants in a group of large sites.

In POD, the research team mailed enrollment packets to potential participants that contained the informed consent form, baseline survey, and information about current-law rules and the benefit offset being tested (Hock et al. 2020). The enrollment packets instructed beneficiaries who wanted to enroll to return the signed consent form and baseline survey. The demonstration's call center staff were available to answer questions, but no telephone or in-person contact was required to enroll in POD. A total of 6 percent responded to the mailing, by returning the enrollment packet; but 2.5 percent who responded refused consent, and 1 percent did not pass the intake screening or did not provide complete information (Hock et al. 2020). Altogether, 2.4 percent of recruited SSDI/SSI concurrent beneficiaries enrolled in POD.

Several reasons might explain lower enrollment in BOND and POD compared to the youth demonstrations (PROMISE, YTD) and to AB, MHTS, and SED. For BOND, the $1-for-$2 benefit offset applied only to earnings that exceed an annualized level of Substantial Gainful Activity (SGA) after completing the Trial Work Period (TWP) and Grace Period. This offer of a financial incentive triggered by future earnings might seem like a more abstract concept than an offer of tangible services such as health insurance or employment services. It seems plausible that for some beneficiaries, uncertainty over whether they would be able to achieve and sustain the level of earnings needed to take advantage of the offset could have dissuaded them from volunteering for the demonstration. In fact, findings from the BOND process analysis reported by Derr et al. (2015) suggest that some of the beneficiaries who did enroll

[13] We are referring to Stage 2 of BOND, where beneficiaries were recruited to volunteer. In Stage 1, assignment to the treatment group was not voluntary.

expressed some uncertainty about whether they would be able to take advantage of the benefit offset.

The lower enrollment in POD compared to BOND could be influenced by the differences in the earnings rules in the two demonstrations. In POD, monthly earnings exceeding the monthly TWP level triggered the benefit offset, whereas in BOND, benefits were not offset unless earnings exceeded the SGA level. The lower earnings threshold for the POD benefit offset would reduce total income for beneficiaries earning between TWP and SGA compared to current-law rules. In contrast, the benefit offset rules in BOND would not reduce total income for beneficiaries under any earnings scenario. The more appealing earnings rules offered in BOND might have influenced the higher enrollment rates. In addition, it is possible that the differences in recruitment methods might have influenced the higher enrollment in BOND. The follow-up telephone calls, and in-person enrollment sessions used in BOND might have fostered a stronger connection to the demonstration and encouraged higher enrollment.

The enrollment packet mailed to potential participants in POD noted that individuals choosing not to enroll did not need to return the consent form and baseline survey. However, the instructions also noted that all beneficiaries who returned the survey and consent form would receive the $25 incentive payment. Hock et al. (2019) report that early results from the January 2018 outreach mailing showed the number of beneficiaries who returned the materials but declined to consent exceeded the number who returned the forms and consented to enroll. The POD researchers tested several alternative recruitment procedures in a pilot to identify ways to refine and enhance the recruitment process. One change tested was an insert to the initial enrollment packet with clearer instructions that only those who wished to enroll needed to return the consent materials and baseline survey. After making this change, the research team observed a reduction in the number of letters returned by those who declined to enroll. That is, 5.2 percent of the January 2018 mailing sample returned the materials but declined to consent, whereas only 2.1 percent of those sent a February 2018 enrollment packet (with the insert) returned the materials and declined to consent.

The pilot also found that follow-up postcards were as effective at boosting enrollment as follow-up phone calls, and less costly. A postcard sent ahead of the enrollment packet, and a last chance reminder postcard also increased response (Hock et al. 2019; Hock, Wittenburg, et al. 2020). The pilot results provided immediate evidence about ways to tailor the POD recruitment process. In addition, the researchers noted that the insights from the pilot might also have other applications relevant to SSA. In particular, the finding that postcards were as effective as phone calls in spurring response to outreach is potentially useful for other SSA administrative procedures such as letters about TTW and notifications about continuing disability reviews, where phone calls would likely be infeasible or costly.

In collaboration with the Office of Evaluation Sciences (OES) at the General Services Administration, SSA has conducted several experimental impact studies to evaluate alternative messages. Informed by behavioral insights, a recent study evaluated four variations in reminder letters intended to encourage SSI recipients to report changes in earnings (GSA/OES 2019b). The study found that receiving any one of the letters increased earnings reporting and the amount of countable earnings reported, potentially reducing overpayments. Hemmeter et al. (2020) reported the results of another study aimed at improving participation in SSI among individuals over age 65. The researchers found that receiving any one of four letters informing potential applicants about their likely eligibility for the program, emphasizing the simplicity of the application process, or noting the maximum monthly benefit increased SSI applications and awards. SSA also collaborated with OES to evaluate the effect of providing information about employment assistance available at American Job Centers or state Vocational Rehabilitation (VR) agencies to denied applicants (GSA/OES 2019a). The impact analysis found that providing such information had no effect on appeals. SSA is collaborating with OES to design a new study to examine alternative information intended to increase participation in the TTW program.

Using dedicated recruitment staff who do not also have responsibility for service delivery has shown advantages for achieving enrollment results and service delivery goals.

In the previous section we describe the various approaches used in the demonstrations for recruitment; key features of outreach and recruitment in the demonstrations are shown in Exhibit 9.4. Our comparison of recruitment strategies shows several advantages of using dedicated recruiters rather than having the same staff conduct recruitment and service delivery.

The PROMISE projects adopted a customized process for recruitment, modifying the number and type of contacts. Projects sent initial letters and then used texts, calls, email, and in-person visits to encourage enrollment. As reported by McCutcheon et al. (2018), in the New York project, recruitment took extensive effort—with 41 percent of enrollees receiving between 6 and 10 contacts before enrolling, and another 12 percent receiving 11 or more contacts. Local project staff in five of the six PROMISE projects conducted the outreach and recruitment. The Maryland project used a local contractor, dedicated solely to recruitment. It achieved the target sample ahead of schedule and the highest enrollment rate (43 percent) of all the PROMISE projects (Kauff et al. 2018).

Exhibit 9.4. Key Features of Demonstration Outreach and Recruitment

	Uniform Recruitment Procedures Used in All Sites	Research Contractor Conducted Recruitment	Enrollment Required In-Person Meeting	Enrollment Required Additional Eligibility Screening beyond SSA Data	Same Staff Conducted Recruitment and Delivered Services
AB	✓	✓		✓	
BOND	✓	✓	✓		
MHTS		Contractor and demonstration staff	✓	✓	
POD	✓	✓			
Project NetWork			✓		✓
PROMISE[a]			Varied		In three programs (CA, NY, WI)
SED	✓	✓	✓	✓	
TETD			✓	✓	✓
YTD	✓	✓			

Source: Authors' summary of demonstration reports. AB: Michalopoulos et al. (2011). BOND: Gubits et al. (2018a/b). MHTS: Frey et al. (2011). POD: Hock et al. (2020). Project NetWork: Kornfeld and Rupp (2000). PROMISE: Anderson et al. (2018); Honeycutt, Gionfriddo, Kauff, et al. (2018); Kauff et al. (2018); Mamun et al. (2019); Matulewicz, Katz, et al. (2018); McCutcheon et al. (2018); Selekman et al. (2018). SED: Taylor et al. (2020). TETD: Thornton and Decker (1989). YTD: Fraker, Mamun, et al. (2014).
[a] The Maryland PROMISE project engaged with a local contractor to conduct recruitment and enrollment.

In AB, BOND, POD, SED, and YTD, the research contractor conducted outreach and recruitment. It sent initial mailings, made follow-up phone calls, and conducted informed consent and random assignment. To reinforce the legitimacy of the outreach and promote trust, the contractor worked closely with SSA to develop outreach materials and messages and to format letters to make it clear that the contractor was contacting the individual on behalf of SSA.

Focusing exclusively on recruitment (rather than balancing recruitment with service delivery) and applying techniques used in survey data collection—tracking all contacts, calling at different times of day, obtaining additional contact information—to maximize response rates could have contributed to the results for AB, BOND, SED, and YTD. Where the project staff were responsible for both recruitment and service provision—in Project NetWork and two of the PROMISE projects (CA, NY), the process analyses reported challenges in balancing the two functions, with delays in service provision during peak recruitment times (see Leiter, Wood, and Bell 1997; Matulewicz, Katz, et al. 2018; McCutcheon et al. 2018).

Furthermore, the PROMISE projects in which the same staff were responsible for recruitment and for service delivery also had some advantages. This approach helped staff build rapport and trust with the youth during recruitment, which helped to encourage the youth to participate in services after enrollment. The process analysis

also found that the continuity in staffing avoided disruption that could come with a handoff to a different staff person after enrollment. In New York, this staffing arrangement impeded service delivery. Early on it was necessary to focus intensively on achieving enrollment targets, making it challenging to give the attention needed to engage enrolled youth in case management at the same time. This staffing arrangement could have led to delays in beginning case management with enrolled youth in New York, where the average time from enrollment to first contact was reportedly 220 days (McCutcheon et al. 2018).

The sites that had staff focused solely on recruitment appear to have completed recruitment more quickly. They also used similar approaches—mailings, with follow-up calls and postcards, and in-person meetings.

Additionally, the individual approach to recruitment seems to have yielded better results than conducting group sessions. As reported by Honeycutt, Gionfriddo, Kauff, et al. (2018) and Matulewicz, Katz, et al. (2018), the Arkansas and New York projects initially tried group sessions as the first step in recruitment (initial letters invited youth to a group meeting), then changed to individual contacts when they were unable to recruit enough youth using only the group sessions.

Early interventions face tradeoffs between achieving the target sample size and conducting adequate screening to identify the intended target population.

Early interventions such as DMIE, sponsored by the Centers for Medicare and Medicaid Services, and RETAIN, sponsored by SSA and the US Department of Labor (DOL), seek to provide services to workers at risk of leaving employment because of illness or injury. Intervening early, before workers separate from the labor force and before they apply for disability benefits is an important policy priority. Services provided soon after an onset of illness or injury might be more effective at preserving employment than interventions that begin later (Ben-Shalom, Christian, and Stapleton 2018). Recruitment and enrollment for these interventions can be challenging because of a lack of information about which workers are at risk of leaving the labor force and of applying for disability benefits and which of them could benefit from an early intervention. Identifying the desired population is essential to target resources efficiently. Another challenge is determining how and where to identify such workers.

As Anderson et al. 2020 point out, to maximize the likelihood of detecting effects of an early intervention such as RETAIN, it is crucial for programs to consider the tradeoffs between achieving the target sample size and adequate screening to identify the appropriate target group. The ideal target group is workers who would leave the labor force and enter SSDI or SSI in the absence of an intervention. This group is difficult to identify—many workers who experience an illness or injury will remain in the labor force even if they do not receive any assistance other than health care.

Unless the demonstration offers the intervention to a target population who would be likely to be eligible for SSDI, without enrolling a large sample it will be difficult to detect effects of the intervention on SSDI, if such effects exist. Imagine if the AB

recruitment process had not screened on health insurance, and instead randomly assigned all new SSDI beneficiaries to receive its health insurance and employment services. That only 11 percent were uninsured would have created treatment and control groups consisting largely of beneficiaries who had health insurance, potentially jeopardizing the ability to evaluate the effects of removing the 24-month waiting period. The additional screening of the new SSDI beneficiaries was crucial to identifying the desired target group for the evaluation and to establishing the extent to which newly entitled SSDI beneficiaries had unmet health needs during the 24-month waiting period.

Anderson et al. (2020) point to the experience in DMIE and data on the geographic variation in SSDI application and awards as a lesson for RETAIN. In DMIE, the impact analysis did not detect impacts of the health care and employment services on SSDI application and found that only a small proportion of the control group lost a job or applied for disability benefits. This indicates that recruitment and enrollment in DMIE might not have recruited a sample of the workers most likely to benefit from its interventions. Given the finding in DMIE, Anderson et al. suggest that for RETAIN, a customized recruitment approach in each state will likely be more successful than attempts for federal sponsors to standardize recruitment. The states could need to use large catchment areas and solicit workers from the entire state to achieve the desired sample size after applying the screening needed to identify the ideal target population.

Comparing Volunteers to Non-Volunteers

Even if the sample is large enough to support analyses, it is also crucial to examine who enrolls and how closely the sample compares to the target population of interest. The composition of the group of volunteers determines whether the findings from the sample would be applicable if the policy were offered more broadly. The composition of the volunteers can also affect the likelihood that an evaluation can detect intervention impacts. This section highlights two lessons that arise from the analyses of volunteers versus non-volunteers reported in the demonstrations' evaluations.

Outreach to a broad group of disability beneficiaries produced volunteers who are distinct from the general caseload in their orientation toward work.

The broad target groups for BOND, POD, and Project NetWork, which were unconditioned on type of disability or other factors, could be readily identified in SSA administrative data without additional eligibility screening. Researchers and policymakers expected that the financial incentives offered in the demonstrations would attract beneficiaries with an interest in work. The benefit offset offered in BOND and POD and the Project NetWork waiver that stopped the TWP for a period of 12 months would only be advantageous to beneficiaries who expected to earn at a level where the incentive would take effect.

In all three demonstrations, the analyses showed that the beneficiaries who volunteered appeared more inclined to work than did non-volunteers. For example, Project NetWork found that volunteers were more likely than non-volunteers to have worked 30 hours per week in the past year, and volunteers reported more positive attitudes and commitment to work than non-volunteers did. Volunteers also were less likely to report poor health and less likely than non-volunteers to report a limitation that prevented work.

The BOND Stage 2 volunteers had higher rates of employment at baseline compared to the nationally representative Stage 1 group, with 36 percent of the Stage 2 control group working in the year prior to random assignment, compared with 14 percent of the Stage 1 control group (Gubits et al. 2018a/b). The BOND evaluation found that Stage 2 volunteers were more likely to be women, were younger, and more likely to have a mental health disorder than were non-volunteers (Gubits et al. 2013). Beneficiaries who received SSDI benefits for 36 months or less volunteered at higher rates than those with longer SSDI receipt, and disabled adult children were less likely to volunteer, as were those who had a representative payee appointed to help manage benefits.[14] As Livermore (2011) found in an analysis of 2004 National Beneficiary Survey data, work-oriented beneficiaries were more likely to be younger and shorter-duration beneficiaries.

The POD evaluation also found that compared to non-volunteers, a higher proportion of volunteers had a history of substantial earnings, including a higher proportion with recent history of earnings at the TWP level, at the SGA level, or between TWP and SGA. In addition, compared to non-volunteers, a higher proportion of POD volunteers had engaged with a TTW Employment Network (Hock et al. 2020). Overall, the POD researchers concluded that patterns of differences between volunteers and non-volunteers are consistent with past research about factors that differentiate work-oriented beneficiaries in the SSDI caseload. This suggests that like Project NetWork's and BOND's processes, the POD recruitment process produced a sample of beneficiaries with greater orientation toward work than in the full caseload.

In YTD, the researchers concluded that the enrollment process yielded a broad group of youth SSI recipients who were like non-enrollees. As expected for the youth target population, the volunteers were not more likely than non-volunteers to have had recent work experience or higher earnings in the previous year. In PROMISE, the projects found that volunteers were slightly younger than non-volunteers. Like all the demonstrations, in both YTD and PROMISE, that the group of volunteers are self-selected likely means that volunteers differ from non-volunteers in unobservable characteristics such as motivation or interest. In the process analyses for PROMISE, authors (Anderson et al. 2018; Honeycutt, Gionfriddo, Kauff, et al. 2018; Kauff,

[14] The Social Security Act authorizes SSA to appoint representative payees if it determines that program beneficiaries/recipients are unable to manage their own benefit payments. Beneficiaries with a representative payee were less likely to volunteer for BOND and POD compared to beneficiaries without a representative payee.

Honeycutt, Katz, et al. 2018; Mamun et al. 2019; Matulewicz, Katz, et al. 2018; McCutcheon et al. 2018; Selekman et al. 2018) caution policymakers that impact results are not likely generalizable to the full sample of youth SSI recipients but are indicative of results for a group who would volunteer for the package of services offered.

Findings about which beneficiary characteristics are associated with enrollment could help SSA target recruitment efforts in the future.

In addition to reporting the proportion of those eligible who enroll and comparing the characteristics of volunteers to non-volunteers, several demonstrations conducted more rigorous analyses of participation patterns (see Burstein, Roberts, and Wood [1999] and a summary by Ruiz-Quintanilla et al. [2006]). Project NetWork examined participation rates among subgroups defined by program and personal characteristics. Its analysis found the highest participation (12.2 percent, compared to the overall 4.5 percent enrollment rate) among those who had worked more than 30 hours per week in a job in the 12 months prior to enrollment, who did not report severe limitations in activities of daily living, and who reported they were able to work.

Heckman and Smith (2004) decomposed participation in DOL's Job Training Partnership Act experiment into stages: eligibility, awareness, application, acceptance, and enrollment, modeling characteristics associated with each stage. Using a similar approach for one of the SPI demonstration projects, Ruiz-Quintanilla et al. (2006) examined four stages of participation in the New York WORKS project: (1) information delivery; (2) response; (3) interest; and (4) enrollment. At the first stage, delivery refers to an informational letter not being returned as undeliverable because the person no longer lived at that address. The researchers examined the relationship between individual characteristics and the outcomes at each stage in the recruitment process. For example, at the last stage, the researchers analyzed the effect of individual characteristics on enrollment given that the person was eligible, the letter was not returned, the person responded, and expressed interest in their response. The results showed that younger SSI recipients were more likely to not participate, because they did not respond to the letter. The results also showed that SSI recipients with anxiety disorders who expressed an interest in the project were more likely to drop out at the enrollment stage than were SSI recipients with other psychiatric disorders.

Building off this approach, both MHTS and SED also conducted analysis to predict the factors related to enrollment (see Frey et al. [2011] and Taylor et al. [2020]). In addition to showing the potential for engaging with individuals before SSDI entitlement, the results from SED (reported in Taylor et al. [2020]) indicate that certain characteristics affect enrollment. Compared to non-enrollees, they found that men, those with less prior work experience, and those with higher educational attainment were more likely to enroll. Local context also influenced enrollment, with those living in areas with higher unemployment and in counties where average wages were rising more likely to enroll. Another predictor of enrollment was denial at step 5

in the disability determination process,[15] indicating a decision by disability adjudicators that the individual was unable to perform the same work as in the past ("past relevant work") but could perform alternative work in the national economy. This finding might suggest a potential target population for future policy tests. The recruitment analysis also examined reasons for not enrolling and found that concerns about the legitimacy of the demonstration offer was a common concern, both among enrollees and non-enrollees. Other reasons for not enrolling included the perception that the potential volunteers could not work or improve their health.

As reported by Frey et al. (2011), the MHTS found that several items available in SSA administrative data predicted enrollment: having a representative payee, distance from the study site, months receiving SSDI, and recent TTW activity. The researchers concluded that enrollment might exceed 25 percent in a demonstration if SSA were to target SSDI beneficiaries with recent TTW activity. This suggests that offering a specialized, intensive service like what was offered in MHTS might be particularly attractive to beneficiaries who have shown an interest in employment services. The main reasons for declining to participate in MHTS included general lack of interest, concerns about not being able to work, and concerns about physical health.

In the next section we explore what happens after recruiting the demonstration sample to identify lessons about implementing the various types of interventions SSA has tested in its demonstrations.

LESSONS ABOUT IMPLEMENTING AN INTERVENTION

To identify lessons about implementing interventions, we limited our review to demonstrations where process analysis findings are available. We include lessons from experimental and non-experimental designs, from implementing services in the local and state demonstrations Benefit Offset Pilot Demonstration (BOPD); Homeless Outreach Projects and Evaluation (HOPE); Homeless with Schizophrenia Presumptive Disability (HSPD) Pilot; SSI/SSDI Outreach, Access, and Recovery (SOAR); and SPI, as well as in the large national experiments AB, BOND, MHTS, POD, Project

[15] SSA evaluates disability applications in a five-step determination process: (1) The SSA field office determines whether an applicant is financially eligible for SSDI or SSI. (2) If so, a Disability Determination Services (DDS) examiner evaluates whether the applicant has a severe impairment; those without a severe impairment are denied benefits. (3) DDS examiners determine whether the applicant's mental or physical impairment meets or medically equals an impairment in the Listing of Impairments; those that do, result in an award. (4) For those that do not, the DDS evaluates whether the applicant's residual functional capacity is sufficient to perform past relevant work or (5) whether the applicant can perform other work in the national economy. For more detail see the publicly available Program Operations Manual System*:* https://secure.ssa.gov/poms.nsf/lnx/0422001001. POMS DI220001.001 discusses sequential evaluation of Title II and Title XVI adult disability claims.

NetWork, PROMISE, Structured Training and Employment Transitional Services (STETS), TETD, and YTD.

We first highlight the contributions of SSA's earliest demonstrations in showing the feasibility of recruiting applicants and beneficiaries and delivering interventions that promote employment. We then organize the discussion into two groups: (1) lessons from demonstrations that evaluate changes to SSDI program rules in the form of benefit offsets; and (2) lessons from demonstrations that evaluate specialized services outside of SSA operations.

SSA's earliest demonstrations laid the groundwork for future research and contributed to the body of evidence about supported employment. SSA's TETD followed the DOL's STETS conducted in the early 1980s; both showed that it was feasible to recruit participants with intellectual disability and to deliver employment services involving direct placement in competitive jobs (Kerachsky et al. 1985; Thornton and Decker 1989). The STETS demonstration was one of the first evaluations of transitional employment for youth with disabilities. SSA subsequently began TETD in 1985 to evaluate transitional employment services for youth and adult SSI recipients with intellectual disability. SSA awarded grants to eight private, non-profit, and university-based organizations to provide services in 13 sites, including job development and placement, on-the-job training, and short-term follow-up support. Some of the organizations had experience providing these services; others created new programs for TETD.

TETD placed two-thirds of the treatment group members in jobs, and half of them were stabilized in permanent jobs. As reported by Thornton and Decker (1988) and Decker and Thornton (1995), the basic program elements in TETD were implemented as planned, with some variation across sites, but all sites delivered the essential components. The implementation findings showed that 12 months seemed to be an adequate amount of time to find and place participants in permanent jobs, and that a wide variety of supportive services (e.g., job search assistance, soft skills training, housing, and budgeting assistance) were necessary to respond to the diverse needs of the target population. It also showed the critical nature of transportation assistance for employment support and the extensive efforts needed to assist participants.

Project NetWork showed that it was possible to recruit a broad cross section of the SSDI beneficiary, SSI recipient, and applicant populations. In the mid-1990s when it began, Project NetWork was the largest demonstration SSA had conducted, with outreach to 145,404 potential volunteers. Project NetWork tested the effects of case management provision on employment. In one model, SSA field office staff delivered the case management; in another, SSA field office staff implemented the less intensive referral management. Private rehabilitation organizations delivered case management in the private contractor model, and state VR agencies implemented a model where state VR counselors provided case management from an SSA field office. The case managers coordinated the rehabilitation process; obtained medical, psychological, and

vocational assessments; established vocational goals and plans; and monitored participants' progress.

The process study showed that 60 percent of participants completed assessment and employment planning, and 45 percent received purchased employment-related services across all four of the case management models (Leiter, Wood, and Bell 1997). Kornfeld and Rupp (2000) concluded that broad-based return-to-work services can be implemented on a large scale in a variety of institutional arrangements. Project NetWork was an immediate precursor to the TTW program, authorized in the Ticket to Work and Work Incentives Improvement Act of 1999 (Ticket Act).

SPI showed it was feasible to support states to implement innovative strategies to promote employment for SSI recipients and SSDI beneficiaries. In 1998**,** SSA funded 12 of the state projects and the US Department of Education's Rehabilitation Services Agency (RSA) funded six. Specific components varied, but the projects provided services in these general areas: (1) improving information about the effect of work on benefit receipt (benefits counseling), (2) encouraging the use of available work incentives, (3) testing modifications to program rules to allow SSI recipients to earn and save more, and (4) providing better access to vocational supports.

Despite mixed results from the impact analyses conducted in four of the SPI projects, the conclusions report discusses several ways that SPI informed future program design (Kregel 2006a, 31–32):

- The SPI projects led the way to establish a nationwide system of Benefits Planning, Assistance, and Outreach (BPAO) projects, with many staff involved in the ongoing training provided to these BPAO projects. The BPAO projects became the WIPA program that currently provides benefits counseling through SSA.
- Several SPI projects were instrumental in facilitating the development and/or implementation of Medicaid buy-ins in state projects, at first through the Balanced Budget Act of 1997 and later through the Ticket Act.
- The model for the Disability Navigators initiative within the One-Stop Career Center system that is currently under DOL's Employment and Training Administration was initially developed through the RSA-funded Colorado SPI project.
- In several SPI projects, the use of benefits planning and assistance services by the state VR became a "routine" component of service delivery for SSA beneficiaries.

These early demonstrations highlighted the array of services that can be offered and the range of organizations that could collaborate to offer employment services to SSDI beneficiaries and SSI recipients. They also contributed to the landscape of disability research and helped set the stage for SSA's ongoing TTW program.

Lessons from Implementing Benefit Offsets

Implementing interventions that change SSDI earnings rules have posed unique implementation challenges.

In BOPD, BOND, and POD, SSA evaluated earnings rules that adjust SSDI benefits through a benefit offset in place of the current law's "cash cliff."[16] These demonstrations have evaluated changes to the SSDI program rules, as well as changes to the processes for beneficiaries to report earnings and for SSA to adjust benefits. Implementing these changes has posed challenges and offers some lessons.

Benefit Offset Pilot Demonstration (BOPD). SSA conducted BOPD to generate lessons for implementing the national offset demonstration in BOND. (The pilot also produced initial estimates of the likely impact of the benefit offset for volunteers.) As reported by Chambless et al. (2011) and Tremblay et al. (2011), one of the most important lessons from BOPD had to do with administering the benefit offset. In BOPD, SSA used a manual process to calculate benefit payments according to the demonstration rules. SSA customized this process for each of the four participating states. Though this manual process minimized disruption to SSA's current-law operations for processing earnings and calculating benefits, it created delays in adjusting benefits. Also, some beneficiaries received notices with incorrect information about their SSDI benefits; and in some cases, errors applying the offset rules led to under- and overpayments to beneficiaries. To handle the much larger size of the national demonstration and improve on the implementation experience in the pilot, BOPD recommended that for BOND, SSA automate and centralize the administrative procedures used to adjust benefits.

Benefit Offset National Demonstration (BOND). Drawing on lessons from BOPD, SSA developed an automated benefit processing data system to calculate benefits in BOND (Gubits et al. 2013; Gubits et al. 2018a/b; Stapleton et al. 2010). This system operated separately from SSA's regular administrative data systems to avoid disruptions to current-law operations. In another step to avoid disrupting regular operations, SSA attempted to approximate the benefit offset implementation that would occur in an ongoing program, but without involving the SSA field office structure. Therefore, it established a centralized team at SSA to assist with the administration of the BOND case processing and contractor staff to obtain earnings estimates, document earnings deductions, and assist with SSA notices and appeals.

Even with the automated benefit adjustment system and centralized SSA operations team, Gubits et al. (2018a/b) report that timely benefit adjustment was

[16] Under current-law program rules, SSDI beneficiaries lose all SSDI benefits after substantial earnings. This complete loss of benefits is referred to as the "cash cliff." Specifically, SSDI benefits are suspended or terminated if, after completing a nine-month TWP and a three-month Grace Period, a beneficiary's countable monthly earnings exceed the monthly SGA amount.

challenging in BOND. Median duration from first month of offset use (defined as the first month when a beneficiary's earnings triggered the benefit offset) to the time that SSA first adjusted SSDI benefits was 22 months for Stage 1 and 15 months for Stage 2. For Stage 2, enhanced counseling led to shorter times to first benefit adjustment compared to standard work incentives counseling, most likely because of the proactive outreach of enhanced work incentives counseling staff, which in turn might have improved beneficiary reporting.

Delayed first adjustments meant that beneficiaries continued to receive full SSDI benefits after the time that benefits should have been reduced by the benefit offset. Delays in benefit adjustment could have diminished beneficiary understanding of and confidence in the offset rules. As Gubits et al. (2018a/b) report, factors that contributed to delayed benefit adjustment included backlogs in conducting the work continuing disability reviews necessary to determine when the TWP had been completed and when benefit offset should be applied and that some beneficiaries were not timely in reporting earnings.

As discussed by Derr et al. (2015), the initial notifications to the Stage 1 treatment group explained the benefit offset and provided contact information for the demonstration's call center and website, but the notifications did not direct beneficiaries to contact the demonstration staff. Information provided to the Stage 2 volunteers during recruitment and after assignment to the treatment group (Gubits et al. 2013) provided instructions for reporting earnings and a message about the importance of timely reporting of earnings. More consistent and clearer messages about the requirement to report earnings and procedures for doing so might have improved earnings reporting.

Another lesson from BOND relates to challenges in replicating the level of knowledge of new earnings rules as would occur in a national program. The levels of understanding of BOND rules in the Stage 1 and Stage 2 research samples (Gubits et al. 2018a/b) suggest that outreach and information were not sufficient for the treatment group to understand the offset rules as well as the control group understood the current-law rules. (That knowledge of current-law earnings rules is itself low; in BOND, 54 percent of the nationally representative Stage 1 control group seemed to understand them.)

Outreach to the national sample assigned to the Stage 1 treatment group (the mandatory sample) consisted of two letters and two phone contact attempts from the contractor and a notice from SSA. These efforts led to 29 percent of the Stage 1 treatment group, three years after random assignment, knowing correctly how earnings affect benefits under the offset rules (Gubits et al. 2018a/b). The volunteers in the Stage 2 treatment group received outreach and recruitment materials about the benefit offset and completed an informed consent and enrollment process. These extra efforts in Stage 2 produced wider, though still limited, understanding of offset rules among the Stage 2 treatment group compared to Stage 1. About half of the members in each of the two Stage 2 treatment groups correctly understood the offset. Enhancements to

work incentives counseling evaluated in one of the Stage 2 treatment groups increased the level of understanding compared with the other treatment group's understanding (52 percent in the enhanced work incentives group understood the BOND rules correctly, compared to 48 percent in the standard work incentives group; Gubits et al. 2018a/b).

The evaluators concluded that (1) implementation challenges could have been one of four factors that kept those offered the offset from using it.[17] The other three factors were (2) limited work capacity among beneficiaries; (3) insufficient increase in the incentive to earn more; and (4) complexity of the intervention and of the current-law rules, making it difficult for beneficiaries to understand the change in incentive.

The BOND evaluators noted that it is possible that the impact on the proportion of beneficiaries earning more than the BOND threshold might have been somewhat larger in the nationally representative Stage 1 had outreach to Stage 1 treatment group members been more robust and benefit adjustments quicker. Nothing in the evidence, however, suggested to evaluators that the overall finding for BOND would have changed if these implementation challenges had been avoided.

Promoting Opportunity Demonstration (POD). One year after volunteering to enroll in the demonstration, compared to the BOND Stage 2 voluntary treatment group, three times as many POD treatment group members had used its benefit offset (Levere, Mann, and Wittenburg 2020). Compared to BOND, the lower earnings threshold for the benefit offset in POD, different earnings rules, and the lower volunteer rate likely contributed to higher rates of offset use in POD sooner after random assignment. In addition to the benefit offset, POD evaluated alternative rules for several work incentives—eliminating SGA, the TWP, and Extended Period of Eligibility. Another change compared to BOND is that SSA adjusted benefits using monthly earnings rather than annual earnings estimates. These changes reduced SSA's administrative burden for adjusting benefits, relative to BOND where a work continuing disability review was required to determine when the benefit offset should be applied.

In another contrast to BOND, POD distributed benefits processing throughout seven of SSA's payment centers, creating small workgroups in each payment center responsible for any manual workloads necessary to implement the POD rules. POD also offered beneficiaries an online option for reporting earnings that was not available in BOND. Beneficiaries could mail pay stubs, enter the information into the online tool, or report earnings by phone to their POD counselor. This online tool simplified earnings reporting for some beneficiaries. Mamun et al. (2021) report that in the first two years of implementation, 24 percent of POD treatment group subjects had used

[17] In Stage 1 of BOND, 3.6 percent of the treatment group used the offset in any of the first five years after enrollment. As expected, given that a higher proportion of Stage 2 volunteers were working at enrollment, Stage 2 offset use was higher than in Stage 1, with about 15 percent of treatment group members using the offset in any year during the first five years (Gubits et al. 2018a/b).

the benefit offset. Altogether, 22 percent of POD treatment group subjects reported earnings, and the online portal was the most frequent method used (46 percent of those who reported earnings used the online portal).

Work incentives counseling was important for explaining the benefit offset, and findings from BOND showed it was feasible to implement changes in the current counseling model.

The results of BOPD (see Chambless et al. 2011; Tremblay et al. 2011) showed the importance of work incentives counseling, and both BOND and POD included such counseling. BOND also implemented an enhanced form of work incentives counseling that featured increased outreach and intensity of services such as proactive, regular outreach from the counselor, structured vocational assessments, and an employment support plan. BOND assigned smaller caseloads per counselor for enhanced work incentives counseling, and it used performance benchmarks for participant engagement to communicate expectations and monitor progress. The BOND process analysis (see Derr et al. 2015) found that the enhanced counseling was implemented according to design.

Relative to the regular work incentives counseling, Gubits et al. (2018a/b) report that the enhancements yielded positive effects on some important outcomes: improvements in beneficiaries' understanding of the benefit offset rules; shorter average duration from first offset use to benefit adjustment; and lower average overpayments. However, the counseling enhancements did not increase use of the offset; generate higher earnings; or reduce SSDI benefits. Nor did the evaluation find any evidence that the enhancements improved beneficiaries' lives in other areas such as health status, health insurance coverage, participation in other income assistance programs, or household income.

Lessons about Service Delivery

SSA has also evaluated a variety of specialized services that are not part of SSA's regular operations. In this section we explore lessons about implementing these types of services. Drawing on findings reported by Frey et al. (2011), Michalopoulos et al. (2011), Fraker, Mamun, et al. (2014), and Mamun et al. (2019), we found three general approaches to implementing services. First is the approach used in MHTS and SED. These demonstrations evaluated a highly structured intervention implemented in multiple, local program settings. Second is the approach used in AB, in which a single, centralized provider delivered a uniform set of services to demonstration participants in multiple locations. Third is the approach used in PROMISE, RETAIN, and YTD, in which SSA (and its federal partners in PROMISE and RETAIN) established guidelines but gave flexibility to local projects to develop specific service delivery approaches. The lessons highlight factors that contribute to successful implementation and factors that make it challenging to deliver services in each of these arrangements.

When the objective is to produce evidence about the effects of a highly structured, specialized service delivered by local programs, careful attention to site selection and rigorous fidelity monitoring can help ensure the intervention is implemented according to the intended design.

MHTS, conducted from 2006 to 2010, was the first time the IPS form of supported employment was evaluated with SSDI beneficiaries in community-based mental health systems. SED is also evaluating IPS, for denied SSDI and concurrent applicants with mental health conditions. The process study for MHTS showed that the intervention can be delivered with high fidelity to the evidence-based IPS model on a large scale in real-world health care settings. As noted by Bond, Drake, and Becker (2012):

> Individual Placement and Support (IPS) is a systematic approach to helping people with severe mental illness achieve competitive employment. It is based on eight principles: eligibility based on client choice, focus on competitive employment, integration of mental health and employment services, attention to client preferences, work incentives planning, rapid job search, systematic job development, and individualized job supports. Systematic reviews have concluded that IPS is an evidence-based practice. (32)

To avoid uncertainty about the findings—Would a lack of effects indicate that the service does not work, or would poor implementation mean that the demonstration did not perform a reliable test of the intervention?—the research team opted to select a purposive sample of community mental health centers with experience operating the IPS model. They recommended 20 sites of 50 in operation in 2006 when the demonstration began (Frey et al. 2011). SSA also sought to ensure adequate representation of the Hispanic population, and the research team added two additional sites that had experience serving that subgroup. Because the study covered a wide range of geographic areas and included sites that served high proportions of Hispanic beneficiaries, it provided support for the hypothesis that these kinds of services could be replicated in other regions of the United States.

The researchers examined program-level fidelity, the extent to which the programs adhered to IPS standards, using a validated 15-point Fidelity Scale that rates each program on staffing, organization, and service requirements. Researchers conducted annual site visits and rated each program according to the scale (see Frey et al. 2011, Appendix 5A). The study found that 77 percent of the MHTS program sites achieved high fidelity to the IPS model in the first year of the program and even more (86 percent of programs) in the second and third years.

The investigators concluded that this sustained, high level of fidelity was unusual, better than that attained in the National Implementing Evidence-Based Practices Project, which set out to use a comprehensive and standardized training strategy for IPS (McHugo et al. 2007). They attributed the high level of program-level fidelity to

careful site selection and rigorous monitoring. The researchers selected sites purposively to maximize the potential for high fidelity to and consistent implementation of the service model. The result is that the study provides strong evidence about the feasibility of implementing IPS and the impacts of the IPS intervention but less information about whether the findings are generalizable to a national policy. As Barnow and Greenberg note in Chapter 2 in this volume, MHTS is an example of an efficacy trial, providing insight about the optimum implementation of a given intervention.

As reported by Frey et al. (2011), the researchers also examined individual-level service utilization data to determine the extent to which treatment group members engaged with IPS. They found low rates of employer contact by beneficiaries who were not employed, and relatively low rates of receipt of mental health case management services. The demonstration showed low job-seeking rates among the treatment group members who were not employed, and the study was not able to isolate the precise barriers to employment for this population.

The investigators also assigned global ratings of IPS fidelity to each site based on annual assessments derived from a structured checklist that supplemented information from the IPS Fidelity Scale with information about site-level activities and specific requirements for the MHTS implementation. These results showed that 86 percent of the sites had adequate to very strong implementation in the first year and 74 percent in the second year. The most frequently cited barriers to implementation were unresponsive leadership, finances, mental health services not available, and staff turnover.

The authors note that previous research has shown program-level fidelity to be a strong and consistent predictor of participant outcomes (Bond, Becker, and Drake 2011). However, the investigators found no association in the MHTS between program-level IPS fidelity measures and site-level employment rates. It might not have been possible to detect variation in impacts correlated with fidelity simply because there was insufficient variation in fidelity among programs (Frey et al. 2011). The MHTS findings of a lack of correlation between program-level fidelity and treatment group outcomes suggest a need for further research, perhaps even formal evaluations that systematically vary aspects of the implementation, to better understand the extent to which program-level fidelity influences outcomes.

Given policy interest in whether IPS can be an effective intervention for individuals with other types of health conditions, it is important to know whether strict fidelity to the IPS model leads to better outcomes and whether positive outcomes are possible even with less stringent implementation of the model’s components.

The AB demonstration showed that when the goal is to provide a limited set of uniform services in numerous locations, a centralized service provider can be a practical solution to promote consistency and efficiency.

The AB demonstration evaluated whether providing health insurance and employment supports to new SSDI beneficiaries in the Medicare waiting period would improve health and employment outcomes. AB operated in the 53 metropolitan statistical areas in the United States that had the largest populations of new SSDI beneficiaries. The demonstration targeted new SSDI beneficiaries in the Medicare waiting period who did not have health insurance. One treatment group received health insurance; a second treatment group received health insurance plus services intended to promote employment.

The AB Plus treatment group received the health plan and three other voluntary services: (1) Progressive Goal Attainment Program (PGAP), (2) employment and benefits counseling, and (3) medical case management. PGAP is a behavioral modification program intended to incrementally increase activity levels and change daily routines to be consistent with employment. PGAP had not been evaluated in randomized clinical trials prior to AB, but evidence from non-experimental research supported its potential effectiveness for SSDI beneficiaries (Michalopoulos et al. 2011). The demonstration implemented the AB Plus services in a centralized manner, by in which a single service provider organization conducted intake and PGAP, another provider offered employment and benefits counseling, and a third performed medical case management. This approach minimized concerns about variation in service delivery across sites and provided an efficient solution to deliver services across many widely dispersed locations.

PROMISE, RETAIN, and YTD feature federal guidelines for services but offer local programs flexibility to customize intervention design. This approach offers insights about the use of existing services to evaluate a new program design.

In contrast to AB and MHTS, where the services evaluated were precisely specified, the federal sponsors of PROMISE[18] and YTD established guidelines about required services but gave the programs flexibility to design and deliver services. Both demonstrations are based on effective youth transition practices documented in the National Collaborative on Workforce and Disability for Youth's *Guideposts for Success* and the National Technical Assistance Center on Transition's Effective

[18] SSA and its partners (US Departments of Education, Health and Human Services, and Labor) set requirements for core components of the PROMISE projects: (1) formal partnerships with state social service agencies; (2) case management; (3) benefits counseling and financial education; (4) career and work-based learning experiences; and (5) parent training and education. Each project also had to enroll 2,000 participants.

Practices and Predictors Matrix (Fraker, Mamun, et al. 2014). Both demonstrations intended to deliver intensive case management, benefits counseling, financial literacy training, and career and work-based learning experiences.

For RETAIN, DOL established seven core service components,[19] but the states that implement RETAIN have considerable flexibility to design and implement interventions. Each state will develop an approach to identify and recruit its target population and to determine the role of health care providers, employers, and the return-to-work coordinator. As with YTD and PROMISE, process analyses for RETAIN that describe the services delivered in detail, roles of service providers, and implementation results will be essential to understanding the intervention delivered and to interpreting impact estimates.

For YTD, SSA selected universities and private non-profit agencies to deliver the intervention.[20] Except for West Virginia, YTD projects were already serving youth, and some had to modify their focus to deliver the employment-focused services called for in the demonstration. Changing existing program operations and philosophy can be a challenge and required extensive technical assistance in some cases. For example, the Erie County (NY) project adapted its original program model, which used a classroom-based self-determination curriculum along with basic education and career exploration but no direct employment services. To adapt its program to the YTD logic model, the project replaced the classroom-based structure with individualized case management and employment services. The process analysis found that the project delivered a structured set of services that conformed closely to the updated logic model. Alternatively, adapting a prior program model proved more challenging in the Colorado YTD project. As discussed by Fraker et al. (2014) and Fraker, Baird, et al. (2011), a strong commitment to the project's original focus on case management posed a barrier to developing an emphasis on employment services and individualized work-

---

[19] RETAIN's core components are (1) return-to-work coordinators to coordinate health and employment service delivery; (2) training for participating health providers in occupational health best practices and alternatives to opioids for pain management; (3) incentives for participating health care providers to use best practices; (4) early communication to all stakeholders to return the worker to the workplace as soon as possible; (5) workplace-based interventions, including accommodations such as lighter and/or modified duties and adjustments to work schedules, tasks, and the physical worksite, if necessary; (6) training/rehabilitation for workers who can no longer perform their prior job or other available suitable alternate work; and (7) tracking and monitoring the medical and employment progress of participating workers.

[20] As detailed by Fraker, Mamun, et al. (2014), the experimental evaluation included six YTD projects that entered the evaluation in two phases. SSA selected three Phase 1 projects from a group of seven projects SSA had been funding through cooperative agreements, and three Phase 2 projects from a group of five pilot projects that it had been funding through a contract. Phase 1 projects had been operating for several years before the evaluation began, which affected their receptivity to technical assistance. The evaluation found systematic differences between the phases in how the projects were implemented and their impacts on youth outcomes.

based experiences. This strong commitment to the original program model also led to resistance on the part of project managers to technical assistance designed to help staff develop skills to provide job development and job placement services (Fraker, Baird, et al. 2011).

Lessons from YTD influenced the design of the PROMISE demonstration. To improve the study's potential to detect impacts, the PROMISE project sponsors adopted a larger sample size (2,000 per project, compared to 800 per project in YTD). PROMISE also focuses on younger individuals (ages 14–16, compared to 14–25 in YTD) and on serving the entire family. Based on lessons from YTD about the need for coordination across multiple touchpoints—school, health care, Vocational Rehabilitation—PROMISE aimed to secure buy-in and cooperation from all relevant state agencies. One of the YTD lessons, the importance of partnerships for effectively serving youth in transition, became a foundation for PROMISE as it sought to improve the coordination of services.

Examples from PROMISE show that programs that rely on existing relationships might be able to establish partnerships quickly and might have existing services in place that can be tapped for the demonstration. But in some cases, relying on existing service providers taxed their capacity, made it hard for them to commit resources and staff to the new program, and made it difficult to engage in a timely way with program participants (McCutcheon et al. 2018). The ASPIRE (consortium) project provided case management as a new service and then referred participants to existing programs for benefits counseling, financial literacy training, and employment-related services. This proved to be an efficient way to deliver services for this site. The Arkansas project found that developing a new program can be challenging and time-consuming, particularly the work needed to educate community partners. Relying on service providers in different organizations also presented a management challenge for Arkansas.

Relying on existing service providers can also raise concerns for preserving the experimental contrast. Using existing providers to provide demonstration services to the treatment group means that demonstration staff need to ensure that those providers do not also provide demonstration services to the control group. When coupled with low take-up of services by treatment group members, this might make it harder to establish a sufficient treatment/control differential to detect the effects of the PROMISE intended services. Some PROMISE projects, for example, generated greater treatment/control contrast than others and produced a greater chance of detecting impacts. The New York project assigned control group members to receive some contact with case managers who also served the treatment group (McCutcheon et al. 2018). In projects where recruitment staff also provided case management, it was more possible to introduce opportunities for contact with controls. In the Maryland project, a robust existing service delivery system meant that control group members had access to many of the services offered to the treatment group (Kauff, Honeycutt, et al. 2018).

It is important to consider tradeoffs regarding data systems and monitoring for designs that are centralized or locally developed.

As noted in the process analysis reports, each PROMISE project developed its own management information system (MIS) to record information about service delivery, although each PROMISE project was required to use an MIS to record data on recruitment and its efforts with treatment group youth and families. Project differences made it difficult to compare projects on some measures, though that was not an objective for the PROMISE evaluation.

Locally developed systems allow programs to build on existing systems that staff are familiar with, which can be less costly than developing a new MIS and training staff on it. Limitations in some of these systems for measuring service receipt can make it difficult to monitor service receipt in detail, and different MISs might not capture uniform data, which makes it difficult to measure service receipt in the same way across all programs. The PROMISE process reports document some complications with data entry that hindered the ability to measure service take-up. For example, Matulewicz, Katz, et al. (2018) note that in the California project, the staff did not consistently enter data about the enrollment interview, because if a participant enrolled, the interview would be assumed to have occurred. In other cases, some interactions were recorded only in client case notes, making it difficult to identify the service provided. Also, the timing of some events was not recorded, making it difficult to determine whether the project met its benchmark.

The approach taken in other demonstrations to develop a single data system (e.g., BOND, POD, and YTD) was more practical because the organizations that operated the demonstrations were required to build systems under their contracts to SSA. In contrast, the PROMISE projects, as Department of Education grantees, were not required to build data systems. If it were possible, setting uniform content and data entry requirements might have made projects easier to compare. Weighing tradeoffs in costs, deciding on minimum data collection requirements up front, and staff training can help to maximize the value of data systems used to monitor service provision.

Lessons about What Helps or Hinders Service Delivery and Participation

Leadership, mode of delivery, and adjustments during implementation affected service delivery.

The demonstration implementation studies report on several factors that influence service delivery. Here we summarize factors that arose in several demonstrations, affecting participation and engagement.

Leadership. Several factors related to leadership have contributed to successful service delivery: (1) a strong management structure, (2) setting clear expectations and roles across partners for collaboration and communication, and (3) obtaining high-level buy-in for support. PROMISE required grantees to form partnerships among state

departments of education, Vocational Rehabilitation, Medicaid, and other agencies to deliver services. Obtaining high-level buy-in within the state governments in the sites helped to coordinate across the various agencies and to achieve the strong partnerships that were essential to implementing the programs.

The Wisconsin PROMISE project formed an executive committee with top state government officials and a steering committee with agency and provider leaders as part of the management structure to cultivate the partnerships. The committee was important for ongoing communication, coordination of multiple partners, and meeting enrollment targets (Selekman et al. 2018). The Maryland project built on its existing relationships with the state agencies that participated in the project, finding that creating a small leadership team was effective for project operations and communication between partners. The smaller team allowed for clearer messaging to providers and quicker responses for guidance when needed (Kauff, Honeycutt, et al. 2018).

In multiple PROMISE sites, evaluators identified the importance of clear roles and expectations among partners for achieving more effective collaboration at the local level. In the ASPIRE project, which consisted of a consortium of states, the process analysis report notes that a centralized management structure and advisory committee, such as those described above, might have been helpful in facilitating linkages, consistency, and communication across partners (Anderson et al. 2018). To maintain these local collaborations, the California project found that management monitoring progress was a key factor for these ongoing relationships (Matulewicz, Katz, et al. 2018). In the Arkansas site, where multiple organizations provided services, managers faced difficulties setting the expectations and clarifying roles among partners. The site overcame these issues through joint trainings and meetings, including joint management team meetings (Honeycutt, Gionfriddo, Kauff, et al. 2018).

Four other projects are examples where effective leadership has produced collaboration with third parties to improve the disability application process for underserved groups. This lesson is particularly relevant now as SSA responds to growing concerns about the effects of the COVID-19 pandemic on disability applications.

In the early 1990s, SSA conducted the SSI Outreach Demonstration, providing cooperative agreements to outside organizations to conduct outreach and application assistance. Building from lessons from that effort, the HOPE demonstration engaged programs to conduct outreach to individuals experiencing chronic homelessness to help them file disability applications, along with assistance accessing other treatment and services. SSA provided the organizations with trainings and information about the agency's application process. Participants in HOPE who received application assistance received determinations more quickly and had improved housing situations 12 months after enrollment (McCoy et al. 2007). The evaluators identified two important factors for implementation: (1) coordinated collaboration among HOPE

staff, Disability Determination Services (DDS) liaisons, and SSA regional office or field office liaisons; and (2) clear and open communication among these parties, with all invested in problem-solving.

The Substance Abuse and Mental Health Services Administration's SOAR program built off HOPE and was designed to help individuals at risk of or experiencing homelessness to access disability benefits. A SOAR Technical Assistance Center provides ongoing trainings and coordination for state and local initiatives. A central component of the SOAR model is the collaboration among the SOAR sites, SSA, and DDS, as SOAR staff help eligible individuals navigate the disability application process. In an evaluation of SOAR in six states, Kauff et al. (2009) identified strong leadership to facilitate coordination and communication among partners; agency-level buy-in and support; active collaboration; engagement from SSA and DDS; and strong, ongoing, and structured communication among partners as keys to successful implementation. Kauff, Clary, and Lyskawa (2014) found that using core components of the SOAR process for filing SSDI and SSI applications was predictive of higher approval rates at the initial application level. They found the rate was almost double the rate for all applicants experiencing homelessness in Fiscal Year 2010. The authors also found that the collaboration among SOAR programs, SSA, and DDS was crucial.

SSA also conducted the HSPD Pilot to address barriers to receiving SSI benefits for individuals with schizophrenia or schizoaffective disorders. The intervention was based on a structured collaboration among community health agencies, SSA, and DDS. Bailey, Goetz Engler, and Hemmeter (2016) found favorable outcomes for individuals who received assistance, such as higher approval rates at initial application and reduced time to award. The SSA regional office championed this pilot and was invested in the effective collaboration and communication with the community partners.

Mode of Delivery. Demonstrations have provided lessons about the advantages and disadvantages of delivering services in new ways—telephone versus in person, group versus individual benefits counseling. In AB, the demonstration provided the AB Plus services—PGAP, benefits and employment counseling, and medical case management—by telephone because the number of participants in each location was too small to deliver in-person services.

The AB process study found that it was possible to deliver services by telephone but with some limitations. For example, some employment counselors reported challenges in ascertaining a beneficiary's work limitations and work readiness. One counselor helped a participant explore the physical demands of jobs requiring standing or lifting by explaining on-the-job activities. Because the remote counselors could not have direct experience with the labor markets in 53 different metropolitan areas, they would encourage participants to explore their neighborhoods to identify businesses and potential employers. Counselors also developed resource lists with employment services in each area and contacted service providers to collect information on procedures for serving SSDI beneficiaries.

PGAP, one of the services offered to the AB Plus treatment group, was originally designed as a face-to-face intervention for Canadian workers' compensation claimants, delivered by occupational and physical therapists. In AB, social workers provided PGAP, and the demonstration participants had a wider range of diagnoses and functional limitations than previous service recipients. This was also the first time that PGAP was delivered by telephone; before AB, it had been provided only in person. Michalopoulos et al. (2011) reported that overall, PGAP was delivered as intended, but noted that other studies of PGAP that provided this service in person had larger effects than in AB. Moreover, participant take-up was low, with about one-third of the AB Plus treatment group engaged in this service, and low literacy was a barrier that made participating in PGAP more difficult.

We also note the importance of remote services, as many providers have had to pivot to provide various modes of delivery because of the pandemic, which raises additional questions. Is service provision that is local and in person more effective than service provision that is local and remote? Furthermore, is it possible for a non-local provider to approximate the local context given appropriate training and technical assistance? Along with the examples from AB, SSA's operations of the WIPA program provide important insights on these questions.

The WIPA service model emphasizes remote service delivery, and its community work incentives counselors have developed strategies for interacting with beneficiaries via telephone and video conferencing to develop individualized plans and to counsel them on a range of benefits. Some counselors use screen sharing to review documents with beneficiaries. The WIPA program also uses a database that compiles information on the rules of individual state benefits, making it possible for counselors to advise beneficiaries on a wide range of state benefit programs.

For services in group versus individual settings, PROMISE offers lessons about providing benefits counseling in groups. The Arkansas project provided benefits counseling primarily in monthly group training sessions, with individualized benefits counseling reserved for individuals who had questions about their SSI benefits (Honeycutt, Gionfriddo, Kauff, et al. 2018). The Arkansas staff reported that group sessions worked smoothly and were well received by participants, with about half (55 percent) receiving benefits counseling in the group format. However, in the New York PROMISE project, group benefits counseling was not well received, as youth and families did not want to discuss personal financial information in a group. In that project, low take-up of the group benefits counseling prompted the project to move to individual counseling (McCutcheon et al 2018).

Adjustments during Implementation. The process analyses highlight factors that make it difficult to deliver services as planned and opportunities to adjust service delivery to achieve intended goals. MHTS is an example where ethical and practical considerations made it impossible to achieve uniform service implementation for one of the project components, systematic medication management (SMM). Two issues affected implementation of the SMM. First, the demonstration designed an approach

for this service, but it was not practical to expect the providers to change their regular operations. Therefore, some staff began providing SMM according to the demonstration design but also continued serving other individuals as before. In addition, if treatment group members already had an ongoing relationship with a mental health provider who assisted with medication management, they were not required to change providers, and these outside providers were not required to adopt MHTS's practices. It was not considered reasonable or ethical to ask treatment group members to discontinue relationships with existing providers.

In other demonstrations, early assessments identified the need to re-focus services on employment to improve delivery of the intervention as intended. Modifying service approaches and technical assistance made it possible to make these course corrections. For example, in the AB demonstration, as reported by Michalopoulos et al. (2011), early information showed the need for adjustments to ensure services promoted a rehabilitation model rather than a medical model. The process analysis identified adaptations to services that helped to encourage participants to re-orient their health care and daily routines toward returning to work. One change was to remove questions from the AB Plus intake instrument about the individual's medical condition and medications, and instead emphasize preparation for employment-oriented activities. This change came about when, early in the demonstration, the design team became concerned that the original instrument focused too much on medical providers and medications and that this distracted participants from engaging with employment and benefits counseling and PGAP. After the change, most of the intake time focused on introducing PGAP. Another adaptation was to restrict referrals to medical case management to specific short-term medical issues that were limiting a beneficiary's ability to initiate PGAP.

Course corrections and technical assistance in one of the YTD projects also helped to maintain a focus on employment. The projects had been operating prior to YTD and sometimes had to alter prior practices. As discussed by Fraker, Mamun, et al. (2012), the Miami-Dade County project, previously focused on serving in-school youth with case management and pre-employment services, needed to broaden its focus to deliver the YTD services. In the first year, the process analysis found that participants were spending relatively few hours on employment and paid work experiences. Technical assistance helped the project make changes to put greater emphasis on job placement, which increased participation in employment.

The New York PROMISE project experienced a similar issue. The project staff found that they had not clearly communicated expectations about employment-related services. McCutcheon et al. (2018) found that most referrals for employment services were for pre-employment activities—assessments, career planning and preparation activities. Referrals to paid and unpaid employment were much lower and well below benchmarks the project had set. This occurred because the project wanted to tailor services to youth's needs and did not prescribe benchmarks for types of employment

services. When staff realized that paid employment lagged expectations, they developed increasingly detailed benchmarks and began monitoring more closely.

Local resources and appropriate staffing facilitated, whereas emergency and basic needs impeded, participation and engagement.

Our review of process analysis reports found that local resources and appropriate staffing levels can provide advantages for participant engagement. We also found in many demonstrations that participants experienced emergency needs for housing and food assistance and faced crises that interfered with program participation.

Local Resources. Local resources can be beneficial for service delivery and addressing barriers to participation and engagement. PROMISE required a core package of services but allowed each project flexibility to customize specific approaches to delivering services. Honeycutt and Livermore (2018) highlight that state and federal collaboration is not sufficient for these cross-cutting services, that local agencies must also be engaged. Mamun et al. (2019) discuss the variation in PROMISE local environments and implementation, and how that can influence the impacts of the intervention. In addition, local barriers such as transportation, labor market conditions, and service availability also influenced implementation of PROMISE projects.

Importantly, local resources helped projects meet the need for cultural awareness and sensitivity in PROMISE sites. The California project hired staff who reflected the diversity of the local communities, who could speak in participants' preferred languages, and who understood cultural sensitivities. These resources were also valuable in the ASPIRE project engaging with and providing services to American Indian populations. Another example is the YTD project in Bronx County (NY) that hired bilingual parents of youth to serve as liaisons for other program participants. These liaisons met with youth and their families in their homes and provided encouragement to participate in college workshops.

In the Maryland PROMISE project, staff engaged participants in urban areas differently from participants in rural areas. As Kauff, Honeycutt, et al. (2018) note in the process analysis report for that project,

> In communities where other service options are plentiful, as is often the case in cities, program staff must make the case for why the new services are unique and better than existing ones. In rural communities, where existing services may be limited, families may be more receptive to new services, but their geographic dispersion may make service provision challenging. (57)

Projects in rural areas encountered challenges with distance. The ASPIRE PROMISE project began to allow case managers to travel to more remote areas and developed options for families to receive services online and in other remote modes (Anderson et al. 2018). The Wisconsin project provided tablets to youth to make it

easier to stay in touch, but spotty cell coverage sometimes interfered. Because of geographic dispersion, the Wisconsin project helped participants address transportation needs, varying access to services, and other unique needs (Selekman et al. 2018). Staff in the Maryland project also noted that service coordination could occur based on a staff member's familiarity with local resources (Kauff, Honeycutt, et al. 2018).

Another example is SED, where the local project operators are drawing on local resources to help participants address crucial food and housing needs and to obtain legal assistance. The projects have compiled contact information for legal assistance providers who can help formerly incarcerated participants expunge or correct criminal system records or obtain assistance with custody arrangements or other legal matters. To address major, ongoing medical and behavioral health care needs (particularly for individuals living in non-Medicaid-expansion states), each SED site compiled an inventory of local low- or no-cost health clinics, including dental and vision sources, where participants could find needed care.

In the projects focused on assisting individuals experiencing homelessness to apply for benefits, awareness of the local environment and meeting the needs of the participants within the local context were also key factors. In HSPD, staff in the community centers were familiar with the target population, knew how to find individuals in need of assistance, and knew ways to maintain contact during service provision. Moreover, the local partners were invested in the project, including the community centers, DDS, and the SSA regional offices and local field offices, all of which were also important in HOPE and SOAR.

Appropriate Caseloads. Appropriate caseload assignment contributes to effective service delivery and participant engagement. For many of the services provided across the demonstrations, caseload size and resource allocation are key factors in service delivery. If caseloads are too high, service intensity, quality, and accessibility can suffer and can lead to reallocation of staff. In PROMISE, the Wisconsin project counselors found it difficult to serve both youth and families, because the family members essentially increased their caseloads (Selekman et al. 2018). In the New York site, the project tasked case managers and family coaches with recruitment, which took time away from their duties to provide services (McCutcheon et al. 2018). After the California project hired additional staff to balance workloads better, managers noted that the quality of services improved (Matulewicz, Katz, et al. 2018).

Not only do these issues affect the provision and quality of services, but they also affect participant engagement. Participants in the Maryland PROMISE project commented on staffing changes (Kauff, Honeycutt, et al. 2018):

> When changes occurred, these parents and guardians felt that they and their youth had to start 'all over' because it did not seem to them that previous staff had shared case notes, resulting in the new staff lacking critical information about the youth and families. (24)

In the New York project, when caseloads were too high or staff changes delayed services, participants became frustrated with waiting, which could affect their engagement. These delays also frustrated staff who referred participants to providers that lacked capacity, only to have their participants wait (McCutcheon et al. 2018).

These issues occurred in other demonstrations, too, including BOND and Project NetWork. Benefits counselors in BOND were tasked with collecting earnings information and documenting earnings deductions for the offset, a duty different from their usual workload to counsel beneficiaries, which they reported competed with the counseling (Derr et al. 2015). In Project NetWork, when there were delays in obtaining the initial assessments of participants, which also affected other services, participants disengaged during these waiting periods (Leiter, Wood, and Bell 1997).

Emergency and Basic Needs. Participants' emergency and basic needs can hinder program participation across demonstrations. In multiple demonstrations, process analyses found that immediate needs of individuals and families—food, health, housing—had to be resolved before they could engage with employment-related services. Family crises and challenges made it difficult to remain in contact with some families—contact information changed frequently—and crises made it difficult for some families to engage. For projects generally focused on employment outcomes, first addressing basic needs was viewed as conflicting with pursuit of employment goals. Before they could focus on program goals, providers needed flexibility to help address these immediate needs, and participants needed flexibility to increase stability. SED provides an example where addressing these critical needs was included as part of the model (Marrow et al. 2020).

Alternatively, staff in multiple PROMISE projects reported that instability, crises, and basic needs hindered contact with participants and participants' engagement in services. The Wisconsin project reported that working with the entire family unit rather than just an individual also revealed the needs and complexities (Selekman et al. 2018). In that project, the service provider addressed basic needs throughout the program as they would develop. The provider continued to reach out to engage these participants, as well, even after participants had temporarily disengaged from services. In contrast, some service providers might have policies to discontinue outreach to disengaged participants, rather focusing only on engaged participants. Selekman et al. (2018) note that the service model allowed providers flexibility and promoted engagement in services.

The Maryland PROMISE project hired staff whose sole responsibility was to reach out to participants who had never engaged or who disengaged. Of those contacted through these efforts, one-quarter of the participants began to engage or re-engaged (Kauff, Honeycutt, et al. 2018). In the ASPIRE project, when transportation posed challenges to participating in in-person events, project staff began allowing families to view recorded or live trainings online to make their participation more convenient (Anderson et al. 2018).

The service team in the AB demonstration reported that AB Plus participants typically needed to address other basic needs before they could move on to the PGAP and other service components. Michalopoulos et al. (2011) also found that AB participants had high rates of unmet medical needs and had gone without treatment or care. The health coverage provided in AB reduced these unmet medical needs, and participants were less likely than the control group to opt not to seek medical care for financial reasons and less likely to forgo a needed prescription.

CONCLUSION

This chapter examined findings about demonstration implementation to understand successes and challenges in recruiting and enrolling participants and in delivering services. Our analysis produced a set of lessons and observations about the factors that hindered and supported implementation, considerations for replicating interventions, and the way implementation influences how policymakers interpret impacts. In some cases, even when demonstrations have not produced evidence of impacts on earnings and employment, lessons about operations and service delivery emerge that can inform policy.

Looking across these lessons, we draw seven conclusions that we believe are especially important for policymakers. They constitute valuable contributions to the body of knowledge about disability policy and research and help set the stage for future demonstrations.

- ***Response to recruitment varies among a targeted population and by intervention.*** Variation in recruitment and enrollment underscores the diversity in the SSDI and SSI caseloads. Though SSA's demonstrations have successfully recruited both broad and specific target populations, those targeting narrowly defined groups and offering specialized services have achieved the strongest response to outreach. Overall, newly entitled SSDI beneficiaries, denied applicants, and youth have been more likely to volunteer than existing SSDI beneficiaries and SSI recipients. Broad appeals offering financial incentives yielded the lowest enrollment rates of the group solicited. More rigorous analysis of patterns of enrollment like the analysis conducted by Ruiz-Quintanilla et al. (2006) could help SSA better understand the characteristics that affect program participation at the various stages in the recruitment process. This type of analysis, with greater focus on patterns of participation by characteristics such as education, race, and ethnicity, could help to enhance program outreach and to ensure that all subgroups have access to programs. In general, we think a greater emphasis on equity would strengthen future implementation studies.

- ***Beneficiaries who volunteer for offers of financial incentives are more work oriented than non-volunteers.*** Offering a broad group of beneficiaries a change in benefit rules or a program waiver that made higher earnings more attractive yielded volunteers distinct among the general caseload in their orientation toward work. Volunteers for these demonstrations were more likely than non-volunteers to have had recent work experience. This suggests that for policies offering financial incentives, targeting recruitment to individuals with recent work history might be more efficient than broader outreach.
- ***Administrative challenges can diminish the behavioral response.*** In BOND, administrative burden and operational challenges in implementing a benefit offset delayed benefit adjustments for participating beneficiaries. These time lags could have made beneficiaries less likely to respond to the intervention. As POD is testing changes that reduce some of the administrative burden that occurred in BOND, we look forward to the final results to determine whether these changes were effective.
- ***SSA can use its experience evaluating third-party assistance for underserved disability applicants to respond to an immediate policy concern.*** SSA can build from lessons learned in the SSI Outreach Demonstration, HOPE, HSPD, and SOAR projects to address concerns during the COVID pandemic about access to disability benefits for underserved populations. Those projects showed that strong leadership and effective communication foster the kind of structured collaboration needed to help individuals navigate the application process. SSA has responded to the current crisis by launching a national outreach campaign and designating new positions to serve as liaisons to work directly with third parties. It can apply the principles used to engage with third parties on those projects to develop guidance for SSA's regional offices, field offices, and state DDS to engage with third-party providers to assist applicants. SSA could also conduct rigorous evaluations of these efforts to continue to build evidence about how best to engage with and assist these underserved populations.
- ***Participants may have basic needs that must be addressed if they are to engage in services promoting employment.*** Across multiple demonstrations, participants faced immediate needs—housing, food, health, transportation—that needed to be met before they were able to fully engage with employment-related services. For example, several of the PROMISE projects found that youth and their families could not engage in benefits counseling or career and work-based learning until their immediate crises were resolved. This suggests that addressing basic needs should be factored into demonstration design in the future where appropriate, possibly as a necessary service or by allowing service providers the flexibility to focus on critical needs. It is also important

to consider how providers can successfully continue to engage participants during these crises.

- ***It is possible to implement highly structured, evidence-based services with fidelity, but more research is needed on whether adaptations to a highly structured model could also achieve outcomes.*** MHTS provides strong evidence that it is feasible to implement the specialized IPS model of supported employment with strong fidelity. Key to the success was careful site selection and rigorous fidelity monitoring. However, underpowered to detect such findings, the study found no association between program-level fidelity and treatment group employment rates. This suggests that more research might be needed to understand whether less stringent application of IPS (or other interventions) might also achieve desired outcomes.
- ***It is possible to implement a more flexible service design, providing guidelines for service components but allowing local innovation.*** PROMISE and YTD offered projects flexibility to design services, allowing policymakers to build evidence for what services and arrangements are most effective. Local flexibility promotes innovation and takes advantage of local system strengths; but without a strong data system and fidelity measures, it is more challenging to determine exactly which service/arrangement influences effects and how to replicate it. A model offering local flexibility puts a premium on rigorous process analysis to document exactly what is delivered and how. One important lesson is that within the constraints posed by the type of project or contractual vehicle (cooperative agreements versus contracts), policymakers should set requirements for the data elements to be collected in a program's data system to ensure the system records all the data needed to document in detail the services provided and to compare implementation and outcomes across sites.

Taken together, the evidence about SSA's demonstration implementation underscores a high level of success in carrying out credible tests of a wide range of interventions. Overall, process analyses indicate that interventions have been implemented largely according to intended design. Robust process analyses have allowed for adjustments when needed to improve implementation, and the absence of intervention effects in several demonstrations does not appear to stem from implementation challenges. However, more rigorous evaluation designs that evaluate alternative implementation conditions are needed in the future to understand definitively the role of implementation in participant outcomes.

Exhibit 9.5. Summary of Recruitment Results

	Enrollment Target	Solicited in Initial Outreach	Responded to Initial Outreach		Eligible for Demonstration		Enrolled in Demonstration		
			Number	Percentage of Solicited	Number	Percentage of Responded	Number	Percentage of Eligible	Percentage of Solicited
Early Experiences with National Demonstrations									
TETD		13,800	2,404	17.4	N/A	N/A	745	5.4	5.4
Project NetWork[a]		145,404	11,838	8.1	N/A	N/A	6,527	4.5	4.5
New York WORKS		41,431	17,275	41.6	N/A	N/A	900	2.2	2.2
Broad Appeals Offering Financial Incentives									
BOND	12,650	238,070	9,047	3.8	N/A	N/A	12,954	5.4	5.4
POD	10,000	419,481	24,910	5.9	N/A	N/A	10,070	2.4	2.4
Specialized Services Offered to Specific Groups									
AB	2,000	22,612	18,545	82.0	2,049	11.0	2,004	97.8	8.9
SED	3,000	21,003	13,375	63.7	11,307	84.5	3,000	26.5	14.3
MHTS	2,000	57,634	17,642	30.6	15,982	90.6	2,238	14.0	3.9
Interventions for Youth: YTD									
Bronx County, NY	880	4,843	1,412	29.2	N/A	N/A	889	18.4	18.4
Colorado (4 counties)	880	2,968	1,332	44.9	N/A	N/A	880	29.6	29.6
Erie County, NY	880	3,183	1,296	40.7	N/A	N/A	880	27.6	27.6
Miami-Dade County, FL	880	5,573	1,955	35.1	N/A	N/A	880	15.8	15.8
Montgomery County, MD	880	N/A	N/A	N/A	N/A	N/A	840	N/A	N/A
West Virginia (19 counties)	880	5,207	1,930	37.1	N/A	N/A	875	16.8	16.8
Interventions for Youth: PROMISE									
Arkansas	2,000	7,459	N/A	N/A	N/A	N/A	2,000	26.8	26.8
ASPIRE	2,000	9,196	N/A	N/A	N/A	N/A	2,051	22.3	22.3
California	2,000	11,271	N/A	N/A	N/A	N/A	3,273	29.0	29.0
Maryland	2,000	4,644	N/A	N/A	N/A	N/A	2,006	43.2	43.2
New York	2,000	13,393	N/A	N/A	N/A	N/A	2,090	15.6	15.6
Wisconsin	2,000	9,150	N/A	N/A	N/A	N/A	2,024	22.1	22.1

Source: Authors' summary of demonstration final reports. AB: Michalopoulos et al. (2011). BOND: Gubits et al. (2018a/b). MHTS: Frey et al. (2011). POD: Hock et al. (2020). Project NetWork: Kornfeld and Rupp (2000). PROMISE: Anderson et al. (2018); Honeycutt, Gionfriddo, Kauff, et al. (2018); Kauff et al. (2018); Mamun et al. (2019); Matulewicz, Katz, et al. (2018); McCutcheon et al. (2018); Selekman et al. (2018). SED: Taylor et al. (2020). SPI New York WORKS: Ruiz-Quintanilla et al. (2006). TETD: Thornton and Decker (1989). YTD: Fraker, Mamun, et al. (2014).

[a] Project NetWork enrolled a total of 8,248 in the evaluation. Of those, 6,527 were enrolled through the outreach and recruitment process. The remaining 1,721 were new SSI applicants recruited by SSA's claims representatives.

Chapter 9

Comment

David Stapleton
Tree House Economics

Wood and Goetz Engler (in "Lessons from Implementation") deserve a great deal of credit for drafting an extensive and valuable review of implementation experiences from several decades of SSA demonstration projects. They have drawn some valuable lessons, with which I largely agree. I consider the implications of their findings and lessons for strategies designed to optimize the value of Social Security Administration (SSA) demonstrations going forward.

FOCUS ON EARLY INTERVENTIONS

The authors' findings reinforce a view I have held for some time: that SSA employment demonstrations should focus exclusively on testing relatively "early interventions"—that is, interventions designed for people at risk for application, applicants, and new beneficiaries or recipients, rather than those designed specifically for long-term beneficiaries or recipients. As others have suggested, there are important reasons unrelated to implementation to do so: the bulk of post-award work activity starts in the first few years after award (Liu and Stapleton 2011; Ben-Shalom and Stapleton 2015), and return to work becomes more challenging the longer an individual is out of the workforce (e.g., separation from past employers and skill deterioration). The authors' comparison of recruitment yields (Exhibit 9.2) adds a practical argument: it is easier to recruit youth with disabilities, applicants, and new beneficiaries or recipients to participate in rigorous demonstrations than it is to recruit from the broad adult beneficiary/recipient population. Other things equal, this means that the task of evaluating a meritorious early intervention will be less difficult than the task of evaluating an equally meritorious intervention targeted at long-term beneficiaries or recipients.

INITIALLY TARGET THOSE MOST LIKELY TO USE THE INTERVENTION AS INTENDED

As other authors in this volume have pointed out (Gregory and Moffitt in Chapter 4; von Wachter and Goldman in Chapter 7 and its Comment, respectively), until we know that an intervention works well for a group for which we expect it to work well, it makes little sense to test it on others. The recruitment findings reported by Wood and Goetz Engler reinforce this view. They point to evidence that it is easier to recruit from target populations that are likely to use an intervention—assuming it is possible to make meaningful distinctions in advance. Accelerated Benefits (AB) and the Promoting Opportunity Demonstration (POD) are polar opposites in this regard.

AB offered health insurance to uninsured Social Security Disability Insurance (SSDI) awardees—individuals expected to need financing for health care—and 98 percent volunteered. In contrast, POD recruited volunteers from the full SSDI beneficiary population for the test of a change in the earnings rules that, based on evidence from the Benefit Offset National Demonstration (BOND) and other past research, was likely to be attractive to a small minority. Only 2.4 percent volunteered.

Build on Initial Success

SSA demonstrations have already shown the value of this lesson. The authors point to a string of demonstrations that follow this approach to testing interventions that reduce barriers to Supplemental Security Income (SSI) and SSDI entry to people with disabilities in exceptionally vulnerable subgroups. Similarly, Wittenburg and Livermore (Chapter 6) point to SSA interventions for youth that gradually build on initial success; and Goldman points to the value of expanding tests of the Individual Placement and Support (IPS) model of supported employment, which have been found to be successful for individuals with serious mental illness, to other disability populations.

BOND illustrates why SSA *should not* test an intervention on a broad population until favorable results have been found for a narrowly targeted population. SSA did, in fact, test the BOND benefit offset on four target populations for which the offset was expected to have substantial impacts, under the Benefit Offset Pilot Demonstration (BOPD). BOPD was a proof-of-concept test, designed to help SSA learn about operational issues prior to BOND. Although BOPD was not intended to provide preliminary evidence on impacts for target populations likely to use the intervention, it provided the opportunity to do so. Each of its four randomized control trials recruited beneficiaries who had signaled an interest in work via an interaction with a specific state agency. One of the two unfavorable results from BOPD was in the implementation domain, as Wood and Goetz Engler point out: major problems in processing of benefit adjustments that led to mistakes and long delays in the adjustment of benefits. The other unfavorable BOPD findings are based on the impact analysis completed by Weathers and Hemmeter (2011): there was no detectable impact on earnings whereas mean benefits increased. SSA, which was legislatively required to conduct a national study, moved forward with BOND before the BOPD impact findings were available.

The BOND evaluation findings are unfavorable in the same ways that the BOPD findings were: problems with the implementation of the benefit offset, despite efforts to fix the issues identified in BOPD; and, as Gregory and Moffitt (Chapter 4 in this volume) point out, impacts on mean earnings and mean benefits were unfavorable in the same way. Thus, despite an enormous investment, we cannot confidently rule out the possibility that better implementation of the BOND benefit offset would result in much more favorable impacts. It is at least arguable that SSA would have learned more, and saved time and valuable resources, if policymakers had not required the

agency to conduct a national demonstration before having completed more-targeted tests designed to verify that the impact results would be more favorable once the implementation problems were well addressed.

Find High-Quality Implementation Partners

Although obvious, this point is so important that it deserves explicit attention. Other authors in this volume have written about the value of SSA collaboration with other federal agencies, state agencies, and private organizations. Wood and Goetz Engler point to attributes of partners that are important to success, including leadership, strong working relationships among partner organizations, ability to innovate, ability to implement an intervention with fidelity, ability to make midcourse corrections, and ability to support recruitment. SSA's experience provides many examples of the importance of these attributes. It is important for SSA to draw on lessons from the many different approaches that it has taken to engagement with partners over many demonstrations.

Implement AND Test Innovations That Improve Access to the Main Intervention

Wood and Goetz Engler identify two important challenges that may make it difficult for demonstration participants to access the intervention being tested: limits on their ability to access information technology and unmet needs for basic necessities.

The COVID pandemic both accelerated the use of virtual services and heightened awareness of the need to increase access to information technology for the most vulnerable populations. Virtual services have both benefits and costs, as illustrated by the authors' discussion of remote counseling services. One cost is limitations on access for some individuals. SSA demonstrations provide an opportunity to develop and test approaches to improving access. The Promoting Readiness of Minors in SSI demonstration in Wisconsin offers an example: it provided tablet computers and data plans to students in rural areas. As SSA starts to address the advantages and challenges of expanding virtual services in the SSDI and SSI application process, its demonstrations could test various ways of delivering services virtually in the application of earnings rules, benefits counseling, delivery of early employment interventions, and other services.

Wood and Goetz Engler point to several demonstrations in which treatment participants with high unmet needs for housing, food, clothing, transportation, child care, and other necessities did not have the capacity to take advantage of the intervention. Unmet needs are also an impediment to recruitment. Interventions that are designed to help participants temporarily meet their basic needs, so that they can take advantage of the intervention and get to the point where they can take care of basic needs on their own, seem more likely to succeed than those that leave such needs unmet.

A recent randomized control trial of "self-directed" mental health services illustrates this point. Cook et al. (2019) found that over two years, providing young adults with major mental illnesses considerable discretion in the expenditure of funds available for their mental health services resulted in a considerable improvement in mental health (the objective of the intervention), employment and educational attainment when compared to use of the same funds for mental health services only. The self-directed design is a less extreme version of the intervention that Liebman (in his comment on Chapter 5) suggests as a control arm for all SSA demonstrations: a cash stipend equivalent to the cost of the main intervention. It is more akin to the self-directed delivery of personal assistance services that most state Medicaid programs have adopted following the successful Cash and Counseling demonstration (Foster et al. 2003). High unmet needs in a demonstration's target population is an important reason to build self-directed services into the intervention itself.

The impact findings from rigorous tests of social program innovations always grab the headlines. If impact findings are favorable from the perspective of stakeholders, the test is likely to be deemed a "success"; but if not, it may be deemed "a failure." The review and comparative analysis of many decades of SSA demonstrations offered by Wood and Goetz Engler illustrates that this dichotomous assessment ignores the knowledge that can be gained from rigorous implementation evaluations. The lessons learned go beyond implications for the conduct of individual tests to include implications for the design of interventions to be tested and the approach to testing. That is what makes this chapter such an important contribution to this volume.

Chapter 9

Comment

Calvin Johnson
US Department of Housing and Urban Development[21]

Woods and Goetz Engler (in "Lessons from Implementation") present two sets of lessons learned from Social Security Administration (SSA) demonstrations—(1) recruiting and enrolling participants and (2) implementing an intervention. Discussion of these lessons focuses on 12 demonstrations with rigorous evaluation designs and complementary process evaluations. The following sections highlight key lessons presented by the authors in each section, as well as additional consideration for future SSA demonstrations.

RECRUITING AND ENROLLING PARTICIPANTS

The implementation of demonstrations as a tool for evidence building is challenging. Getting enough people to respond to a notification announcing the demonstration is a great challenge. Without sufficient response to a notification, there is no demonstration to implement. This challenge requires us to assess how much we know about the ways in which potential participants for a demonstration understand and perceive the services being offered. Services that are easier to understand may result in higher response than those that are not. And services that are perceived as more desirable will present fewer recruitment challenges than those perceived as less desirable.

The authors describe Accelerated Benefits (AB) as a demonstration with services that are easy to understand and highly desirable among targeted participants. AB recruitment efforts presented the offer of health insurance benefits and health benefits with support services among Social Security Disability Insurance (SSDI) beneficiaries who are otherwise required to wait 24 months for Medicare coverage. Eighty-two (82) percent of SSDI beneficiaries targeted for the AB demonstration responded. Among those who responded, 11 percent reported not having health insurance. The offer of health insurance is easy to grasp, and waiving the health insurance waiting period has plenty of appeal. The combination of easy to understand and a desirable benefit undoubtedly contributed to the high-response/high-enrollment pattern among eligible SSDI beneficiaries. Further, as the authors pointed out, the high initial response rate even among the already insured is perhaps an indication of unmet health care needs among SSDI beneficiaries overall.

---

[21] The views expressed in this chapter are those of the author and do not necessarily represent the views of the Department of Housing and Urban Development or the US federal government.

The authors provide additional discussion about how assumptions about perceived desirability of services might be used to inform future demonstrations. Specifically, the assumed desirability of supportive employment among SSDI and concurrent applicants and beneficiaries enables us to refine our understanding of service desirability on demonstration recruitment efforts. The response and enrollment rates for the Supported Employment Demonstration (SED) and the Mental Health Treatment Study (MHTS) provide some insight to the possible motivation for demonstration participation among those already enrolled for services and those denied for services. The targets for MHTS were existing SSDI beneficiaries with schizophrenia or affective disorders. These beneficiaries were assumed to have lower levels of motivation to participate in a supportive employment demonstration. The working assumption was that SSDI beneficiaries had limited recent workforce experience and were therefore less motivated to participate in a supportive employment program. Conversely, targets for SED were denied SSDI and Supplemental Security Income (SSI) applicants assumed to have more recent workforce exposure and for whom access to supportive employment services is extremely desirable. Unlike AB, roughly 85 percent of SED solicited applicants and 91 percent of solicited MHTS beneficiaries were eligible for their respective demonstration. Compared to AB, a much smaller percentage of SED and MHTS solicited who were eligible enrolled in the demonstration. The rate of enrollment among SED solicited (denied SSDI and SSI applicants) was nearly twice that of MHTS (SSDI beneficiaries) for similar supportive employment services.

These three studies highlight the impact that the desirability of an offer has on response to a solicitation and subsequent recruitment. First, demonstrations that tap into unmet needs might be more desirable and result in high-response/high-enrollment patterns. The response will be high even among persons ineligible for service. SSA might engage ineligible responders to understand the motivation behind their response.

Second, narrowly focused demonstrations targeting existing beneficiaries will likely have a lower response rate than demonstrations targeting denied applicants. Current beneficiaries might be less motivated to respond and enroll in demonstrations for which the services have little appeal; whereas demonstrations targeting denied applicants might have greater appeal, resulting in higher response and enrollment rates. With data from both the SED and MHTS, SSA has sufficient information to estimate predictive models to identify the characteristics of denied applicants and SSDI beneficiaries with severe mental impairments who enroll in a supported employment program. This information will be extremely useful in developing targeted outreach materials designed to increase enrollment among applicants and beneficiaries in these two distinct groups.

The Promoting Readiness of Minors in Supplemental Security Income demonstration (PROMISE) and the Youth Transition Demonstration (YTD) provide insight to the use of target outreach materials designed to increase enrollment. Specifically, recruitment staff used mail correspondence followed by phone call and

intense follow-up engagement to consent participants and assign them to intervention groups. Given the high level of motivation among youth toward employment, independent living, and education, the use of mail correspondence followed by phone call, text, and additional mail correspondence proved sufficient to achieve acceptable levels of enrollment. The addition of stakeholder engagement among community groups working with youth likely enhanced the outreach effort and provided additional supports in recruiting targeted youth.

POD illustrated the importance of testing outreach materials to increase response and enrollment. Specifically, providing clearer instructions and follow-up by postcard appeared to have resulted in increased response and consent to enroll. In fact, the use of "last chance" postcards was equally effective as a follow-up phone call in boosting enrollment. SSA has conducted messaging experiments with the General Services Administration's Office of Evaluation Science that demonstrate the utility of using behavioral-informed messages to modify behavior. Continuing to work with the Office of Evaluation Science in the design and testing of messages illustrating the benefit(s) to demonstration enrollment will provide additional options for packaging outreach materials. Additionally, developing follow-up messages focusing on intentional next steps and testing their effectiveness in moving potential participants through the enrollment process might offer additional insight into ways to improve outreach materials as well as make enrollment procedures more efficient.

Finally, projects that use the same team members to perform recruitment and service delivery are challenging. The authors suggest that programs with dedicated staff for each of the two roles have higher response and enrollment rates. Although dual-role staff might do a better job building rapport and trust, staff conducting recruitment and service delivery run the risk of sharing information with control group participants that would be unavailable absent the demonstration. Further, sites with separate recruitment and service delivery staff appear to have more efficient recruitment processes and uniform recruitment efforts across sites. SSA might consider including language requiring separate staff to perform recruitment and service delivery in future solicitations.

IMPLEMENTING AN INTERVENTION

For the "Lessons about Implementing an Intervention" section of the chapter, the authors limited their review to demonstrations with process evaluation findings. Below are comments on a selection from each of the section's three subsections (1) "Lessons from Implementing Benefit Offsets"; (2) "Lessons about Service Delivery"; and (3) "Lessons about What Helps or Hinders Service Delivery and Participation."

Lessons from Implementing Benefit Offsets

Whenever there is a rule change for a demonstration that affects benefit amounts, systems designed to compute the benefit must be updated for participants of the

demonstration. Unfortunately, systems designed to calculate benefits are designed on legacy platforms that do not easily accommodate the application of benefit rule changes for the small number of beneficiaries enrolled in demonstrations. Therefore, SSA has either performed these benefit calculations manually across sites or centralized the function to support larger-scale national demonstrations. Regardless of the approach, demonstrations requiring recalculation of benefits often experience delays in benefit adjustment. These delays also contribute to beneficiaries not fully understanding their new benefit rules, and this in turn can erode their confidence in the new benefit rule.

Lessons learned from the Benefit Offset National Demonstration (BOND) provide examples of the challenges associated with demonstrations that affect benefit amounts. Specifically, BOND established a standalone system to make benefit adjustment calculations without interfering with the existing systems used for benefit calculation. SSA established a centralized team that would implement the benefit offset without burdening SSA field office staff who would otherwise have benefit offset case processing responsibilities. SSA hired contractors to estimate earnings, document earnings deductions, and assist SSA staff with appeals. Finally, SSA provided enhanced counseling to ensure that beneficiaries received clear instructions on the importance of timely earnings reporting.

Despite the creation of a standalone system, the parallel implementation process, and enhanced counseling, long delays persisted for the period from the first month a beneficiary's benefit made them eligible for the benefit offset to the time SSA first adjusted their SSDI benefits.

Like other federal agencies that adjust benefit rules for demonstrations, SSA will need to assess the tradeoff of new systems designs for benefit calculation and parallel implementation of benefits programs. If a benefit rule change is a key feature of a demonstration, it is imperative that beneficiaries understand the rule change and have confidence that the calculations are being perform properly. The first step to ensuring an understanding of and building confidence in the new benefit rule is to calculate the new benefit rule promptly, accurately, and consistently. Additional steps taken by SSA to ensure that outreach and counseling staff were delivering the new benefit rules in a clear and concise manner and that beneficiaries understood the importance of timely earnings reporting and the impact for not doing so were important implementation features that netted noticeable results in the average length of time for SSDI benefit adjustments.

Lessons about Service Delivery

Data collection is such an integral part of any demonstration. SSA demonstrations are no exception. When a demonstration relies on a single data system for tracking recruitment and service delivery activities as well as monitoring program activities, the implementers have a better chance at consistent data collection across sites, than if each site had its own system. A single data system ensures uniformity in data fields

viewable across sites. Despite uniformity in data entry fields across sites, there will undoubtedly be variation in the use of these field and the quality of information entered in them. Nonetheless, with a single data system, SSA and its contractor will have access to demonstration data that in theory should be consistent across sites. Using this single data source, SSA and its contractor can establish quality control routines for ensuring data quality standards across all points of data entry. The execution of quality control routines will enable SSA and its contractor to assess data quality standards and use information from their assessment to improve data quality. Data quality improvement efforts may include training and technical assistance on data entry, data standards, and data validation. The objective is to minimize variation in data quality across sites and address quality concerns as they emerge. Demonstrations relying on a single data collection system built by the implementers of the demonstration provide more flexibility in implementing quality control routines and follow-on data quality improvement activities.

The Department of Education required PROMISE programs to use an information system to collect program data. Perhaps not surprisingly, programs such as PROMISE typically collect information within existing systems that are familiar to staff. Because implementation also varies across programs, PROMISE experienced challenges compiling comparable data across programs. As such, PROMISE lacked comparable data across programs. Unlike for PROMISE, SSA required implementers of BOND, POD, and YTD to build an information system to capture key program data. In doing so, SSA positioned itself to capture consistent program data, ensure data quality monitoring, and respond to data quality issues with staff training. SSA must consider the tradeoffs of each approach in future demonstrations.

Lessons about What Helps or Hinders Service Delivery and Participation

Many demonstrations are implemented locally. Successful implementation of local demonstrations requires an understanding of the local service delivery ecosystem. Implementers of these demonstrations must be willing to collaborate with local partners to ensure placement of demonstrations within that ecosystem. As the authors point out, local resources include service providers with an understanding of the cultural context within which service provision occurs. These service providers typically understand nuances of service engagement that program participants have with institutions across their local community. As such, local providers are often poised to incorporate their understanding of the local cultural context in ways that might otherwise impede the ability of demonstration implementers to effectively recruit and enroll participants as well as deliver services sensitive to this context.

Lessons learned from SED highlight the value of local services in addressing the needs of beneficiaries in their local environments. In that demonstration, the implementers required access to local resources to ensure appropriate variation in service delivery. This included working with local legal aid groups to remove criminal

records that served as a barrier to program participation. SED implementers also worked to ensure that sites had access to affordable health, vision, and dental care.

Additional lessons learned from PROMISE and YTD showed an intentionality of service delivery within local cultural contexts. Specifically, these programs hired diverse staff or bilingual parents to address language barriers as well as to engage beneficiaries in their preferred language. ASPIRE gave particular attention to the cultural context of service delivery among American Indians. The Tribal Councils and numerous Tribes informed the cultural context of service delivery and ensured the cultural competence of program staff.

SSA has sufficient information from its process evaluations that highlight the importance of local resources and local cultural context in implementing its demonstrations. To ensure that all organizations implementing demonstrations are equipped to draw on their local resources and understand local cultural context, SSA might consider including in its solicitation for demonstrations a requirement to have partners that understand and have a successful track record navigating local service delivery ecosystems. The solicitation should require a declaration of how local cultural context will be addressed by demonstration implementers. Further, demonstration implementers should be required to provide examples of similar efforts implemented during prior demonstrations.

IN CLOSING

The authors present lessons learned from SSA demonstrations with a focus on recruiting and enrolling volunteers and implementing an intervention. Both components are necessary for the successful implementation of a demonstration. Effective recruitment and enrollment of demonstration targets require an offer of services that is easy to understand. Offering desirable services that also tap into unmet needs among targets of a demonstration might also increase enrollment. And demonstrations targeting denied applicants might result in higher enrollment, especially given the potential lack of appeal that an offer of service(s) may have among existing beneficiaries.

Focusing on a few of the lessons learned from implementing an intervention, implementing an intervention that affects the amount of the benefit requires additional consideration for calculating the new benefit rules promptly, consistently, and accurately. Otherwise, beneficiaries might not understand the new benefit rule and/or lose confidence in the new rule. Although designing a standalone system for new benefit calculations and hiring or assigning staff to replicate benefit processing procedures for a demonstration might be appealing, there are tradeoffs to consider mostly associated with costs and additional staff burden.

Somewhat related is the need to ensure the collection of comparable data across sites. Because demonstrations are about comparisons, the collection of comparable data across sites is important. Local systems for data collection often exist but lack design comparability to facilitate uniform data collection. These data comparability

issues present additional tradeoffs that will affect a demonstration's ability to make meaningful comparisons.

Finally, implementing an intervention requires an understanding of the local cultural context within which services will be delivered. The authors highlight relevant examples of how attention to local cultural context and intentionality with respect to staffing and partners for service delivery are important considerations for implementing interventions. To ensure that demonstrations are respectful to local cultural contexts, it is critical that local partners are engaged to inform the design and implementation of service delivery models.

Implementing a demonstration is challenging, but attending to these and other lessons in this chapter will improve the implementation of SSA's future demonstrations. For sure, other federal agencies will benefit from these implementation lessons, too.

Appendix

Demonstration Descriptions

Sarah Prenovitz and Austin Nichols
Abt Associates

This appendix summarizes the demonstrations and related evaluations described elsewhere in this book.[1] Listed alphabetically, each summary briefly describes the demonstration's intervention (the change in policies or programs being evaluated), the intended target population for the intervention, and the evaluation findings, if available.

Accelerated Benefits (AB)

Purpose: The Accelerated Benefits demonstration tested whether providing health insurance to new SSDI beneficiaries would improve their health and earnings outcomes. The demonstration recruited a sample of new SSDI beneficiaries, ages 18–54, without health insurance and with at least 18 months of the Medicare waiting period remaining.

Timing: Enrollment took place from 2007 to 2009, and the demonstration continued until participants had completed the 24-month Medicare waiting period. Reports were released in 2010 and 2011.

Intervention: Participants were randomly assigned to three groups: a treatment group that received health care coverage (the AB group), a second treatment group that received health care coverage as well as care management and benefits counseling (the AB Plus group), and a control group. The AB health plan offered up to $100,000 of coverage and was more generous than Medicare (covered more services, had lower copays, offered higher reimbursements to providers).

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): 53 sites in the metropolitan statistical areas with the largest number of new SSDI beneficiaries.

Number of Participants: 400 AB treatment, 611 AB Plus treatment, 986 control.

[1] *Key:* SSA=Social Security Administration. SSDI=Social Security Disability Insurance. SSI=Supplemental Security Income.

Research Components: Experimental impact analysis; implementation analysis, which included participation studies; and cost analysis (gross costs only).

Impacts: Access to health care coverage through the demonstration increased the use of medical services and decreased out-of-pocket medical costs and unmet care needs. Access to the AB Plus services increased the likelihood of searching for work, but did not further increase health care use and did not increase employment in the year after randomization. Ongoing research will evaluate the effect of AB through 11 years for employment outcomes and 13 years for SSA disability benefit outcomes.

Further Reading: Michalopoulos et al. (2011); Weathers et al. (2010); Weathers and Bailey (2014).

Benefit Offset National Demonstration (BOND)

Purpose: In the Ticket to Work and Work Incentives Improvement Act of 1999, Congress directed SSA to test the effects of a $1 for $2 benefit offset on SSDI beneficiaries' work efforts. BOND's Stage 1 used a nationally representative sample of SSDI beneficiaries younger than age 60. Stage 2 recruited a sample of SSDI beneficiaries who did not also receive SSI benefits who were expected to be most likely to use the offset.

Timing: Enrollment took place from 2011 through 2012. The final evaluation report was issued in 2018 (though some in the treatment group remain eligible for BOND benefit rules until 2022).

Intervention: In Stage 1, subjects were randomly assigned to an offset treatment group (with standard work incentives counseling) or to a current-law control group. The Stage 1 treatment group was subject to the offset benefit rules, which reduced benefits by $1 for each $2 in annual earnings above the annualized Substantial Gainful Activity level after beneficiaries exhausted their Trial Work Period and Grace Period. This replaced the "cash cliff" that SSDI beneficiaries face under current-law rules. In Stage 2, volunteers were randomly assigned to (1) be covered by the BOND rules and receive standard work incentives counseling; (2) be covered by the BOND rules and receive enhanced work incentives counseling; or (3) be in the control group.

Waivers: See "Intervention" above.

Counterfactual Condition: The Stage 1 control group was subject to current-law earnings rules, under which SSDI benefits are reduced to $0 for earnings above the Substantial Gainful Activity level after the Trial Work Period and Grace Period have been exhausted. Analyses compared the Stage 2 enhanced work incentives counseling (EWIC) group versus volunteers who were subject to the BOND earnings rules who received standard work incentives counseling (WIC) and versus beneficiaries subject to the current-law earnings rules (control group).

Location(s): 10 sites: Alabama; Arizona/Southeast California; Colorado/Wyoming; DC Metro area; Greater Detroit, MI; Greater Houston, TX; Northern New England; South Florida; Western New York/Northern Pennsylvania; and Wisconsin.

Number of Participants: Stage 1: 77,101 treatment, 891,429 control. Stage 2: 3,041 EWIC treatment, 4,854 WIC treatment, 4,849 control.

Research Components: Experimental impact analysis, process analysis, participation analysis, and cost-benefit analysis.

Impacts: The BOND evaluation found no evidence of an impact of the benefit offset on average earnings either in the nationally representative Stage 1 or in the Stage 2 sample of volunteers. In contrast, the evaluation found that the benefit offset increased SSDI benefits due in the five-year follow-up period, in both Stage 1 and Stage 2. Eligibility for enhanced benefits counseling increased the use of those services, but did not increase use of the offset, generate higher earnings, or reduce SSDI benefits.

Further Reading: Gubits et al. (2018a/b).

Benefit Offset Pilot Demonstration (BOPD)

Purpose: The BOPD prepared SSA for the national $1 for $2 benefit offset demonstration (Benefit Offset National Demonstration) by testing the administrative procedures involved in operating a benefit offset. The demonstration targeted SSDI-only beneficiaries receiving benefits based on their own work record who were less than 72 months beyond the end of their Trial Work Period.

Timing: Contracts were awarded in 2004 and enrollment occurred from August 2005 to December 2006. The treatment group was covered by the alternate earnings rules for six years following their Trial Work Period.

Intervention: Alternate earnings rules for SSDI benefits, which replaced the "cash cliff" with a $1 reduction in benefit for every $2 in annual earnings above annualized Substantial Gainful Activity.

Waivers: See "Intervention" above.

Counterfactual Condition: Business as usual.

Location(s): Connecticut, Utah, Vermont, and Wisconsin.

Number of Participants: 923 treatment, 897 control.

Research Components: Experimental impact analysis conducted by SSA by pooling the data for the four states. Each state also conducted its own impact analysis. Process analysis was conducted at the state level.

Impacts: The benefit offset tested in the BOPD led to a 25 percent increase in the proportion of beneficiaries with earnings above the annualized Substantial Gainful Activity amount, had no effect on earnings, and increased benefit payments.

Further Reading: Weathers and Hemmeter (2011); Tremblay et al. (2011); Chambless et al. (2011); Porter et al. (2009); Delin et al. (2010); State of Connecticut (2009).

Benefits Entitlement Services Team (BEST)

Purpose: The Benefits Entitlement Services Team assisted people experiencing chronic homelessness in applying for SSA disability benefits to determine whether it improved timeliness of application processing.

Timing: Implementation ran from 2009 to 2013. Results were published in 2014.

Intervention: Each site included both medical and case management staff, who completed applications, requested existing medical documentation, provided physical and mental health evaluations, and assisted with other tasks such as identifying a representative payee.

Waivers: None.

Counterfactual Condition: NA.

Location(s): Four sites in and around Los Angeles, CA.

Number of Participants: 1,134 initial or reconsideration applications were submitted through BEST.

Research Components: The demonstration was a proof-of-concept study to see whether the project would result in increased program entry and quicker determinations. The non-experimental evaluation compared outcomes of persons served by BEST grantees versus national averages.

Impacts: Allowance rates were substantially higher than the national average (85 percent initial and 90 percent final versus 47 percent initial and 57 percent final for all applications in 2010.) BEST applications also had processing times lower than the national average: 45 days versus 90 days on average during the same period.

Further Reading: Kennedy and King (2014).

Demonstration to Maintain Independence and Employment (DMIE)

Purpose: The Centers for Medicare & Medicaid Services tested whether early medical assistance and employment supports could increase employment and reduce reliance on SSDI or SSI. The demonstration focused on working-age adults who were not yet

qualified to receive federal disability benefits. Each participating state determined its own specific target population. Hawaii focused on a population with diabetes, Minnesota and Texas focused on those with mental health impairments, and Kansas included a variety of subgroups.

Timing: Enrollment and services took place from 2006 to 2009. The final evaluation report was released in 2011.

Intervention: Each participating state designed its own program that included case management, health coverage, and employment services.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): Hawaii, Kansas, Minnesota, and Texas

Number of Participants: Minnesota 1,155; Texas, 1,585; Kansas 500; Hawaii 184; divided between the treatment and control groups.

Research Components: Experimental impact evaluation.

Impacts: The Kansas and Minnesota interventions had modest positive impacts on employment, whereas the Texas and Hawaii interventions did not. None of the interventions discernably affected average earnings.

Further Reading: Whalen et al. (2012).

Homeless Outreach Projects and Evaluation (HOPE)

Purpose: SSA funded third-party outreach and application assistance for homeless and other underserved populations.

Timing: SSA awarded HOPE grants in 2004, and HOPE programs continued to operate through 2009.

Intervention: SSA provided grantees with funding and the *HOPE Program Orientation Manual*, and it convened annual conferences for grantee staff. Grantee organizations conducted outreach to people experiencing homelessness and provided assistance completing applications for SSA disability benefits.

Waivers: None.

Counterfactual Condition: Half of control group organizations received the *HOPE Program Orientation Manual*; the other half did not.

Location(s): 41 programs spread across the United States.

Number of Participants: Data on SSA applications were obtained for 3,055 HOPE participants (about 60 percent of those served) and for 214 applicants served by control agencies.

Research Components: The evaluation included non-experimental impact analyses, focus groups of program administrators, and in-depth site visits of five sites. Impact analyses compared outcomes for people served by HOPE grantee organizations versus those served by similar organizations that did not receive HOPE funding.

Impacts: The evaluation found that people served by agencies that received HOPE funding received SSA determination decisions about a month faster than people who were served by matched comparison agencies. There was no difference in determination time between agencies that received only the *Manual* versus those with no intervention, and no impacts of HOPE on allowance rates.

Further Reading: McCoy et al. (2007).

Homeless with Schizophrenia Presumptive Disability Pilot (HSPD)

Purpose: SSA tested the effect of providing assistance applying for SSI as well as presumptive disability benefits on award timeliness among persons experiencing homelessness who had a confirmed diagnosis of schizophrenia or schizoaffective disorder.

Timing: The pilot was implemented from 2012 to 2014. The report was published in 2016.

Intervention: Community partners provided assistance applying for SSI and recommended presumptive disability benefits while applicants were waiting for a decision.

Waivers: SSA allowed presumptive disability payments (up to nine months of benefits) for the treatment group.

Counterfactual Condition: Business as usual.

Location(s): Three sites in California.

Number of Participants: 260 individuals were served. Analyses are based on a sample of 238 treatment group members, 1,038 individuals from the same site who applied for SSI in the previous two years (C1), 676 individuals who applied for SSI from surrounding areas in the same period as the treatment group (C2), and 857 individuals who established claims in the pilot area in the same period (C3).

Research Components: Non-experimental impact analysis. People served by HSPD were compared with three comparison groups with similar characteristics.

Impacts: The evaluation of the HSPD compared those served by the program versus three control groups with similar characteristics. The intervention increased the allowance rate, decreased requests for consultative examinations, and increased cumulative benefits. Effects on adjudication time varied by the component of the adjudication process and comparison group used. There was no discernable effect on mortality. The fraction in payment status was 23 to 39 percentage points higher after a year in the treatment group.

Further Reading: Bailey, Goetz Engler, and Hemmeter (2016).

Mental Health Treatment Study (MHTS)

Purpose: The Mental Health Treatment Study tested the effectiveness of supported employment and mental health treatment on the employment of SSDI beneficiaries and SSI recipients with severe mental illness. The study recruited a sample of SSDI beneficiaries whose primary impairment was schizophrenia or affective disorder ages 18–55 who lived within 30 miles of a treatment site. Beneficiaries were excluded if they were already working in a competitive job or receiving supported employment services, they had a life-threatening or terminal physical health condition, lived in a nursing home, or had a legal guardian.

Timing: Recruitment began in November 2006, and implementation continued to July 2010. The final report was released in 2011.

Intervention: The treatment group received employment services delivered according to the Individual Placement and Support model. They also received systematic medication management, comprehensive health care, nurse coordinator counseling, and assistance with mental health and return-to-work expenses. The program also covered costs from obtaining services and prescription medications associated with behavioral health care that were not paid for by other sources.

Waivers: Treatment group participants received a three-year suspension of medical continuing disability reviews.

Counterfactual Condition: The control group received a manual detailing local and national supports and services, and $100 in exchange for participating in interviews.

Location(s): 23 sites nationwide except for the Southwest.

Number of Participants: 1,121 treatment, 1,117 control.

Research Components: Experimental impact analysis, implementation analysis of fidelity to the treatment's supported employment model, analysis of gross costs, analysis of utilization of provided services. In addition, scales were used to assess each site's medication management services.

Impacts: Study services increased employment (the treatment group had an employment rate at 24 months of 60.5 percent, compared to 40.3 percent for the control group) and reduced hospitalization rates as of 24 months after randomization. The intervention had no discernable effect on SSDI benefits. The treatment group had statistically significantly higher monthly earnings ($148.16, compared to $97.41 for the control group), statistically significant improvement in mental health status and quality of life, but a slight decline in physical health status. The intervention had no detectable effect on Substantial Gainful Activity: 8.2 percent in the treatment group and 8.8 percent in the control group earned above the SGA threshold.

Further Reading: Frey et al. (2011).

Nudging Timely Wage Reporting: Field Experimental Evidence from the United States Social Supplementary Income Program

Purpose: SSA partnered with the White House's Social and Behavioral Sciences Team to test whether SSI recipients could be nudged to be more timely in reporting their wages, to reduce improper payments. The target population was SSI recipients who were ages 18–50, spoke English as their primary language, were neither institutionalized nor had a representative payee, had been SSI recipients for less than six years, and were somewhat likely to be selected for a continuing disability review.

Timing: Nudging letters were sent on April 15, 2015. Analyses covered calendar year 2015.

Intervention: The study assigned sample members either to a control group or to one of four treatment groups. Each of the treatment groups received a different reminder letter about wage reporting: (1) simple information only, (2) simple information and social information on reporting behavior, (3) simple information and information about the penalties for non-compliance, and (4) all three types of information. The control group received no letter.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): National.

Number of Participants: 50,000.

Research Components: Experimental impact study and cost-effectiveness analysis.

Impacts: Receiving a letter increased the likelihood of reporting earnings and the amount of earnings reported in the three months following receipt of the letter, but the effect decayed over time. There were no differences between the effects of the four messages.

Further Reading: Zhang et al. (2020).

Ohio Direct Referral Demonstration (ODRD)

Purpose: ODRD tests whether direct referrals to Vocational Rehabilitation providers increased Vocational Rehabilitation take-up among youth ages 18–19 receiving or applying for SSI or SSDI.

Timing: Recruitment began in January 2020. The evaluation is expected to continue through December 2022.

Intervention: The Ohio Division of Disability Determination will directly refer members of the treatment group to the Ohio Bureau of Vocational Rehabilitation.

Waivers: Direct referral requires a waiver of existing SSA rules.

Counterfactual Condition: The usual-services group will receive information about Vocational Rehabilitation.

Location(s): Ohio.

Number of Participants: 750 (planned).

Research Components: Experimental impact analysis.

Impacts: Evaluation results have not yet been released.

Further Reading: SSA (2019b).

Project NetWork

Purpose: Project NetWork tested the effects of case management on the employment of people with disabilities. The demonstration targeted SSDI beneficiaries, SSI recipients, and applicants for SSI residing in the areas served by Project NetWork.

Timing: Planning began in 1991, and sites operated during the period from 1992 to 1995. Each site operated for two years, beginning in 1992 or 1993.

Intervention: Services for the treatment group were delivered according to one of four models, with two sites implementing each of the models. In the first three models, treatment subjects met individually with a case or referral manager who arranged for rehabilitation and employment services, helped develop an individual employment plan, and provided direct employment counseling services. Models used various staffing approaches, with one staffed by SSA field office staff, another by private contractors, and a third by state Vocational Rehabilitation counselors working in SSA field offices. The fourth model tested a less intensive referral management intervention delivered by SSA field office staff.

Waivers: For SSDI beneficiaries, waivers exempted earnings for a 12-month period when computing Trial Work Period months and prevented benefit suspension for those who already had exhausted the Trial Work Period. For SSI recipients, the waivers prevented earnings from triggering a medical continuing disability review as would happen under current law.

Counterfactual Condition: Volunteers assigned to the control group received the same waivers of SSDI and SSI rules as the treatment group. The control group could not receive services from Project NetWork but remained eligible for any employment assistance already available in their communities.

Location(s): Eight sites: Dallas, TX; Fort Worth, TX; Phoenix, AZ/Las Vegas, NV; Minneapolis, MN; New Hampshire; Richmond, VA; Tampa, FL; and Spokane, WA/Coeur d'Alene, ID.

Number of Participants: 8,248 volunteers were assigned to either treatment or control status; some analysis of 138,613 eligible nonparticipants at the eight sites.

Research Components: Process study, participation analysis, experimental impact study, and cost-benefit analysis.

Impacts: The Project NetWork services increased treatment group earnings by $220 per year over the first two years following random assignment, but the demonstration had no impact on SSDI or SSI benefit receipt. For the 70 percent of the sample with three-year follow-up data available, there was no impact on earnings in the third year after randomization.

Further Reading: Kornfeld et al. (1999); Kornfeld and Rupp (2000); Rupp, Wood, and Bell (1996).

Promoting Opportunity Demonstration (POD)

Purpose: Section 823 of the Bipartisan Budget Act of 2015 directed SSA to test the effects of a $1 for $2 benefit offset on SSDI beneficiaries' employment outcomes and benefit receipt.

Timing: The demonstration is taking place from 2017 to 2021. Recruitment took place between January 2018 and December 2018.

Intervention: Volunteers were randomly assigned to one of two treatment groups or to a control group. For the treatment groups, POD replaces the SSDI cash cliff and several work incentives policies with a policy that reduces benefits by $1 for every $2 of earnings above the Trial Work Period level (or the amount of Impairment-Related Work Expenses up to the Substantial Gainful Activity threshold). Both treatment groups are subject to the POD earnings rules and receive POD-specific benefits counseling. Volunteers can withdraw from the treatment group and return to current-

law rules at any time. In one treatment group benefit entitlement continues when benefits are reduced to zero because of earnings. In the other treatment group, SSA terminates SSDI entitlement after 12 consecutive months of zero benefits.

Waivers: See "Intervention" above.

Counterfactual Condition: Business as usual.

Location(s): Alabama; Connecticut; Vermont; and parts of California, Maryland, Michigan, Nebraska, and Texas.

Number of Participants: 3,343 treatment 1; 3,357 treatment 2; 3,370 control.

Research Components: Experimental impact analysis, process analysis, participation analysis, and cost-benefit analysis.

Impacts: As of the interim evaluation, which examined outcomes one year after enrollment was complete, POD did not have statistically significant effects on earnings, employment, benefits, or income. However, being in either POD treatment group did increase employment and the likelihood of either being employed or looking for work. A final evaluation report on longer-term impacts will be released in the future.

Further Reading: Hock et al. (2020); Mamun et al. (2021); Wittenburg et al. (2018).

Promoting Readiness of Minors in SSI (PROMISE)

Purpose: SSA and the US Departments of Education, Labor, and Health and Human Services tested whether providing a variety of services to youth and their families improved earnings and employment, reduced reliance on public benefits, and improved other aspects of life. The target group was youths ages 14–16 currently receiving SSI benefits, living in an area covered by a PROMISE site, and not residing in an institution.

Timing: The first sites in the demonstration began enrollment in 2014, and the last ended services in 2019. Final report is due in 2022.

Intervention: Each Education-funded site designed its own intervention program based on federal requirements, including providing four required services: case management, benefits counseling and financial education, career and work-based experiences, and training and other resources for parents. PROMISE placed particular emphasis on encouraging family involvement and creating partnerships between relevant state agencies.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): Arkansas; California; Maryland; New York; Wisconsin; and a consortium of Utah, North Dakota, South Dakota, Montana, Colorado, and Arizona.

Number of Participants: 1,805 to 3,097 in each of the six sites, 13,444 total.

Research Components: Process analysis, experimental impact analysis, and cost analysis. There will later be a cost-benefit analysis.

Impacts: 18 months after randomization, PROMISE had increased youths' receipt of transition services and family receipt of support services at all sites, as well as youth employment. Only one site found a reduction in SSA payments, and four found an increase in total youth earnings income. No impacts were found on youths' expectations, self-determination, or Medicaid use, nor on parental earnings, employment, or income.

Further Reading: Mamun et al. (2019).

Promoting Work through Early Interventions Project (PWEIP)

Purpose: SSA is supporting two existing projects being conducted by the Administration for Children and Families (under the name *Innovative Strategies for Addressing Employment Barriers Portfolio*). Both projects provide services to low-income individuals with limited work histories, to see whether the funded services reduce SSI applications.

Timing: This demonstration is composed of two projects. The Building Evidence on Employment Strategies for Low-Income Families Project (BEES) takes place from 2017 to 2022. The Next Generation of Enhanced Employment Strategies Project (NextGen) takes place from 2018 to 2023.

Intervention: Both BEES and NextGen are examining multiple programs that deliver different interventions in order to generate evidence on their impacts, operations, and costs. Programs include Bridges from School to Work; Families Achieving Success Today; Individual Placement and Support for individuals with justice involvement; Work Success; and Wellness, Comprehensive Assessment, Rehabilitation, and Employment.

Waivers: None.

Counterfactual Condition: Varies by site.

Location(s): Sites include San Diego, CA; Portland, OR; Nashua, NH; Chicago; a regional program headquartered in Louisa, KY; Franklin County, OH; New York, NY; Utah; and Ramsey County, MN.

Number of Participants: To be determined.

Research Components: Experimental impact study, descriptive study, cost study, case study. The two projects encompass up to 22 evaluations, each of which will include one or more of these components.

Impacts: Evaluation results have not yet been released.

Further Reading: Martinson et al. (2021).

Retaining Employment and Talent after Injury/Illness Network (RETAIN)

Purpose: RETAIN is a joint project between SSA and the US Department of Labor testing whether early post-injury/illness health and employment supports increase employment retention and labor force participation and reduce the need for SSDI or SSI benefits.

Timing: RETAIN is taking place in two phases. DOL awarded Phase 1 grants to plan and pilot programs in September 2018, and enrollment in pilot programs began in 2019. DOL awarded Phase 2 grants in 2021 to support broader implementation and more rigorous evaluation, which are expected to end in 2025.

Intervention: State grantees are designing programs modeled on the Centers of Occupational Health & Education (COHE) program. RETAIN programs serve populations of workers who experience injuries or illnesses. They include several key features: (1) training for medical professionals in occupational health best practices, (2) a return-to-work coordinator, (3) efforts to improve communication between the worker, employer, and medical professionals, (4) job accommodations and modifications, and (5) retraining and Vocational Rehabilitation.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): Eight sites participated in Phase 1: California, Connecticut, Kansas, Kentucky, Minnesota, Ohio, Vermont, and Washington. Five states received Phase 2 grants: Kansas, Kentucky, Minnesota, Ohio, and Vermont.

Number of Participants: To be determined.

Research Components: Participation analysis, process analysis, experimental impact analysis, cost-benefit analysis.

Impacts: Evaluation results have not yet been released.

Further Reading: DOL (n.d.).

State Partnership Initiative (SPI)

Purpose: The State Partnership Initiative identified, implemented, and evaluated innovative projects and strategies to provide employment services to SSDI beneficiaries and SSI recipients. SPI included 12 state projects funded by SSA, as well as an additional six projects funded by the Rehabilitation Services Administration (RSA) in the US Department of Education, which are not discussed here. The Employment and Training Administration in the US Department of Labor and the Substance Abuse and Mental Health Services Administration in the US Department of Health and Human Services provided supplemental funding. Projects targeted SSDI beneficiaries, SSI recipients, and people with disabilities more broadly. (Of the 12 projects funded by SSA, 10 reported impact estimates. The six additional projects funded by RSA focused on systems change, so the 18 projects together are sometimes referred to as the *State Partnership Systems Change Initiative*, even though the abbreviation SPI is still used for the combined projects.)

Timing: Funding was awarded in 1998, enrollment began in 1999, and most programs continued through September 2004.

Intervention: Project components included counseling, case management, supported employment, Medicaid buy-in support, and workforce center collaboration.

Waivers: Some of the states implemented waivers to SSI earnings rules, including decreasing the rate at which SSI benefits are reduced for earnings, increasing the amount of unearned income excluded from benefit calculations, allowing higher asset amounts, and suspending continuing disability reviews for SSI-only (non-concurrent) beneficiaries.

Counterfactual Condition: In states that used random assignment in the evaluation to assess the impact of services, the control group was not eligible for those services. As the counterfactual group, the analysis of the effect of SSI waivers used the other SPI states that did not implement waivers.

Location(s): California, Illinois, Iowa, Minnesota, New Hampshire, New Mexico, New York, North Carolina, Ohio, Oklahoma, Vermont, and Wisconsin.

Number of Participants: Most SPI activities were focused on augmenting and changing systems, and as such they did not have a discrete number of participants.

Research Components: Impact analysis, analysis of participation, and implementation analysis. The research design varied by state, with most designs not supporting impact analysis. Three states (New York, New Hampshire, and Oklahoma), implementing four support packages, used experimental designs. An analysis of the SSI Waiver Demonstration Project component of SPI used a non-experimental impact analysis design in which SSI recipients in participating states were compared to SSI recipients in states that did not adopt the waivers.

Impacts: In New Hampshire and Oklahoma, benefits counseling and employment services increased the proportion of beneficiaries who worked during the year after the randomization year by 9 to 17 percentage points. However, in New York, the proportion employed decreased by 30 percentage points. The interventions had either no effect or a negative and statistically significant effect on the earnings of participants, ranging from $1,080 to $1,633 per year.

Further Reading: Kregel (2006b); Peikes et al. (2005).

Structured Training and Employment Transitional Services (STETS)

Purpose: The US Department of Labor funded STETS to test the effects of supportive employment services on employment and earnings for youth with intellectual disability. The demonstration targeted youths ages 18–24 with intellectual disability, with no other impairments, and with limited work experience, and who were receiving SSDI, SSI, or other support from public programs.

Timing: Programs operated from fall 1981 through December 1983. Reports were released in 1985 and 1987.

Intervention: Participants received transitional work services in three phases: (1) training and support in a work environment; (2) on-the-job training; and (3) follow-up support for those working in unsubsidized competitive positions.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): Cincinnati, OH; Los Angeles, CA; New York, NY; St. Paul, MN; and Tucson, AZ.

Number of Participants: 497.

Research Components: Experimental implementation analysis, impact analysis, and cost-benefit analysis.

Impacts: STETS increased earnings and employment as of 22 months after assignment. Employment in the treatment group was 31 percent, compared to 19 percent in the control group. Earnings in the treatment group were $36 per week, compared to $21 per week in the control group. There was no statistically significant change in SSDI or SSI benefit receipt or income.

Further Reading: Kerachsky et al. (1985); Kerachsky and Thornton (1987).

Supported Employment Demonstration (SED)

Purpose: SED tests the effects of supported employment and other services on employment and benefit receipt for denied applicants for SSA disability benefits. The demonstration targets people ages 18–50 who applied to SSDI or SSI on the basis of a mental impairment and were denied benefits at the initial level.

Timing: Enrollment occurred from 2017 to 2019, with services provided for 36 months following enrollment. The evaluation report is expected in 2022.

Intervention: Volunteers were randomly assigned to a control group, a partial-services treatment group, and a full-services treatment group. Treatment group members receive employment services based on the Individual Placement and Support model, as well as medication management and health care coordination.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): 30 sites in California, Colorado, Florida, Illinois, Kansas, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, New York, North Carolina, Ohio, Oklahoma, Oregon, South Carolina, Tennessee, Texas, Washington, and Wisconsin.

Number of Participants: 3,000.

Research Components: Process analysis, participation analysis, experimental impact analysis, and cost-benefit analysis.

Impacts: Evaluation results have not yet been released.

Further Reading: Taylor et al. (2020).

Transitional Employment Training Demonstration (TETD)

Purpose: TETD tested transitional employment services for SSI recipients ages 18–40 with intellectual disability to see whether those services improved employment and earnings and reduced SSI benefit receipt. It was based on the Structured Training and Employment Transitional Services demonstration, a similar project previously fielded by the US Department of Labor.

Timing: Planning began in 1982; enrollment began in 1985; services were provided through 1987.

Intervention: The demonstration staff placed treatment subjects in potentially permanent competitive employment positions that offered on-the-job training. Staff also provided preparation for those jobs in the form of job development and coaching, and they provided or arranged for follow-on support as needed.

Waivers: None.

Counterfactual Condition: Business as usual.

Location(s): SSA provided funding to eight non-profit training organizations in 13 cities to operate the demonstration. Sites were located in Boston, MA; Seattle, WA; Portland, OR; west central Wisconsin; Monmouth County, NJ; Los Angeles, CA; Milwaukee, WI; Chicago, IL; several locations in Pennsylvania (Harrisburg, Lancaster, Philadelphia, Pittsburgh, and York); and Dover, DE.

Number of Participants: 375 treatment, 370 control.

Research Components: Experimental impact analysis, process analysis, informal (i.e., incomplete) cost-benefit analysis.

Impacts: The evaluation of TETD found that in the three-year follow-up period, the intervention increased earnings in each year, increased employment in the third year, and decreased SSI benefits by 2 percent over the three years.

Further Reading: Decker and Thornton (1995); Thornton and Decker (1989); Thornton, Dunstan, and Schore (1988).

Youth Transition Demonstration (YTD)

Purpose: The Youth Transition Demonstration tested various employment supports to increase the employment and earnings and reduce the need for SSDI and SSI benefits among youth ages 14–25 who received or were considered at risk of receiving SSI.

Timing: The first site began enrollment in 2006, and the last ceased operations in 2012. An earlier study not covered here was a precursor to YTD.

Intervention: Sites received SSA funding, technical assistance, and a manual for providing the core service components based on *Guideposts for Success* (NCWD 2019). Each site designed its own program. Services for the treatment group included case management; benefits counseling and financial literacy training; individualized work-based experiences; links to additional supports; family supports; and added social or health services, but varied in intensity across sites.

Waivers: SSA waived certain SSI program rules. The waivers offered a $1 reduction in benefits for every $4 in earnings, extended the student earned income exclusion to youth ages 21 and older, waived benefit cessation if the youth was found to be ineligible at the age 18 redetermination, offered additional opportunities for using a Plan to Achieve Self-Support, and excluded contributions to Individual Development Accounts from SSI calculations.

Counterfactual Condition: Business as usual.

Location(s): Six sites: four counties in Colorado; Miami-Dade County, FL; Montgomery County, MD; several counties in West Virginia; Erie County, NY; and Bronx Borough, NY.

Number of Participants: 5,103, with about 800 per site.

Research Components: Experimental impact analysis, cost study, and process analysis.

Impacts: YTD did not statistically significantly increase the total number of hours of services from all providers. The package of YTD services improved at least one measure of employment outcomes three years after randomization in three of the sites, but not in the other three. Across six programs, earnings were about $200 higher annually (but the overall impact on earnings was not statistically significantly different from zero), employment increased about 4 percent, and disability benefits were more than $500 higher per year, at the end of three years. Unpublished long-term analyses find no substantive impacts on earnings in years three through eight following randomization.

Further Reading: Fraker, Mamun, et al. (2014); Fraker et al. (2018).

Glossary

benefit offset: A reduction in benefits when an SSDI beneficiary's earnings from work increase, as opposed to the current-law rules involving the Trial Work Period (TWP) and cash cliff thereafter.

cash cliff: The complete loss of benefits when earnings exceed the Substantial Gainful Activity (SGA) amount during the Extended Period of Eligibility (EPE) or thereafter. Sometimes referred to as the "benefits cliff."

cost-benefit analysis: A method to weigh an intervention's costs and benefits to various parties (such as beneficiaries, the government, or society as a whole).

counterfactual: Representation of what would happen (counter to the factual condition observed) to a treatment group, absent an intervention. Also known as the "control condition" (for an experimental evaluation) or "comparison condition" (for a quasi-experimental evaluation).

demonstration: A temporary intervention or a package of interventions of limited scale (i.e., not rolled out nationwide or to all beneficiaries or recipients), implemented for the purpose of being evaluated. The information generated by a demonstration can inform decisions about whether the tested change (or some version of it) should be implemented permanently or more broadly. For example, BOND, MHTS, POD, Project NetWork, RETAIN, SPI, and YTD are all demonstrations.

early intervention: Interventions that occur prior to SSDI or SSI award.

evaluation: An evaluation generates the information by which a demonstration can inform decisions about whether the tested intervention (or some version of it) should be implemented permanently or more broadly. Evaluation comes in several varieties, such as summative (measuring *impact*) or formative, and experimental or quasi-experimental.

experimental evaluation: An evaluation design that uses random assignment to assess impacts; also known as a "randomized control trial" (RCT). Non-experimental research designs that seek to assess impacts are known as "quasi-experimental" evaluations.

Extended Period of Eligibility (EPE): A consecutive 36-month period that follows the Trial Work Period and Grace Period. During the EPE, an SSDI beneficiary may still receive payments depending on how much they earn. SSA can pay disability benefits during the EPE if (1) a beneficiary's condition is still disabling and (2) a beneficiary's work is not Substantial Gainful Activity. Disability benefits will end if work constitutes Substantial Gainful Activity after the end of the EPE.

external validity: The ability of an evaluation's findings to be generalized to times, settings, and populations other than those studied; also known as the evaluation's "generalizability."

formative evaluation: Any evaluation that takes place before or during a program's implementation with the aim of improving the program's design and performance. Formative evaluation complements *summative evaluation* and is essential for trying to understand why a program works or doesn't. Results of formative evaluation are incorporated into the program to improve program implementation.

heterogeneity: Diversity, particularly with respect to impacts.

impact: The change in an outcome that is attributable to (caused by) an intervention.

income effect: An economic concept that captures the increase in consumption of a good or service when income or wealth increases, holding all prices constant; this good can include time away from work (also called "leisure" in economics). The income effect predicts people will work less when wages rise, whereas the *substitution effect* predicts people will work more when wages rise. The "net change in work" is the sum of these two effects.

internal validity: An evaluation design's ability to support causal claims (threats to internal validity are those factors that provide alternative explanations other than the intervention for estimated impacts).

intervention: A policy or program change intended to affect participant outcomes; an intervention may or may not be evaluated. Also called a "treatment" by analogy to medicine.

Old-Age, Survivors, and Disability Insurance (OASDI): The programs under the Social Security Act that pay for (1) monthly benefits to retired workers and their spouses and children and to survivors of deceased insured workers (OASI) and (2) monthly benefits to disabled workers and their spouses and children and for rehabilitation services provided to the disabled (DI).

outcome study: An evaluation where the main objective is to see how outcomes can be successfully measured, rather than to estimate impacts as in a summative evaluation.

outcome: The construct that a given program aims to achieve (e.g., employment), either as the ultimate goal of the program or as step toward the program's ultimate goal. How progress toward an end result is measured in a variable.

population-representativeness: The condition in which a sample (such as participants in an experimental evaluation) is representative of a population (such as all SSDI beneficiaries) because the sample is chosen randomly, possibly in strata, so that on average the characteristics of the sample and the population are the same. Population-representativeness is typically required for external validity.

process evaluation: A type of *formative evaluation*, also called a "process study" or an "implementation study," designed to document the implementation details of a demonstration, including recruitment procedures, services delivered as part of the intervention and in the counterfactual control condition, and outcome measurement.

program: A legislatively authorized system of rules and services; for example, Ticket to Work (TTW) is a program. Compare with *demonstration.*

proof-of-concept study: An evaluation where the main objective is to see whether an intervention (the treatment) can be successfully implemented, rather than to estimate impacts as in a summative evaluation.

propensity score method: An approach to study design that involves matching or weighting, using the propensity to receive the intervention, to create a comparison group for a quasi-experimental evaluation.

quasi-experimental evaluation: A type of summative evaluation that assesses the impact of a program by approximating the result that could be obtained via random assignment. Examples include pretest-posttest comparisons of mean outcomes, posttest-only comparisons of mean outcomes, comparison group designs, differences in difference, interrupted time series, comparative interrupted time series, instrumental variables, and regression discontinuity designs. Each faces different threats to validity that do not apply to random assignment designs.

random assignment: A mechanism—a coin toss, roll of the dice, lottery—used in an experimental evaluation to create treatment and control groups such that the two groups are on average alike in all ways with the exception of access to the intervention being tested.

randomized control trial (RCT): See *experimental evaluation*.

Substantial Gainful Activity (SGA): SSA evaluates work activity if someone is applying for or receiving disability benefits under SSDI, or applying for benefits because of a disability (other than blindness) under SSI. Under both programs, SSA generally uses earnings guidelines ("monthly SGA amount") to evaluate work activity to decide whether the work is substantial, and whether SSA may consider a person disabled under the law. For statutorily blind individuals, the monthly SGA amount for 2021 is $2,190. For non-blind individuals, the monthly SGA amount for 2021 is $1,310. Earning above these amounts indicates that work constitutes SGA.

substitution effect: An economic concept that captures the increase in consumption of a good or service when a price drops, holding income or wealth constant so that current consumption is still just feasible; this good can include time away from work (also called "leisure" in economics). The substitution effect predicts people will work more when wages rise, whereas the *income effect* predicts people will work less when wages rise. The "net change in work" is the sum of these two effects.

summative evaluation: An evaluation design that considers the end results of a demonstration (or program), including outcomes and impacts and sometimes the intervention's cost implications via cost-benefit analysis.

Social Security Disability Insurance (SSDI): A federal program that provides benefits to disabled or blind persons who are insured by workers' contributions to the Social Security trust fund. These contributions are based on earnings (or earnings of a spouse or parents) as required by the Federal Insurance Contributions Act (FICA). Title II of the Social Security Act authorizes SSDI benefits. Dependents may also be eligible for benefits from an adult's earnings record.

Supplemental Security Income (SSI): A federal program for low-income aged, blind, and disabled individuals who meet income and resource requirements. Beginning in 1974, SSI replaced the former federal and state programs of Old-Age Assistance, Aid to the Blind, and Aid to the Permanently and Totally Disabled. SSI is funded by general tax revenues, not Social Security taxes.

Ticket to Work (TTW): A program for SSI recipients or SSDI beneficiaries who want to work and participate in planning their employment. Participation in the TTW program increases available choices when obtaining employment services, Vocational Rehabilitation services, and other support services an individual may need to get or keep a job. Participation in Ticket to Work is voluntary and free to SSI recipients and SSDI beneficiaries. When someone participates in the TTW program, they are said to be "using their ticket." The program participant may not be subject to a continuing disability review while using a ticket. TTW is not a demonstration, but it has been evaluated (using a quasi-experimental design).

Trial Work Period (TWP): Allows SSDI beneficiaries to test their ability to work or run a business for at least nine (9) months and still receive full SSDI benefits if they report work activity and their impairment does not improve.

Vocational Rehabilitation (VR): A public program administered by a state VR agency in each state or US territory to help people with physical or mental disabilities become gainfully employed.

youth: A person or people between the ages of 14 and 25.

References

Abraham, Katharine G., and Melissa S. Kearney. 2020. "Explaining the Decline in the US Employment-to-Population Ratio: A Review of the Evidence." *Journal of Economic Literature* 58 (3): 585–643.

Administration for Community Living. 2020. "Community Integrated Health Networks." https://acl.gov/sites/default/files/common/BA_roundtable_workgroup_paper_2020-03-01-v3.pdf.

Aizer, Anna, Nora E. Gordon, and Melissa S. Kearney. 2013. *Exploring the Growth of the Child SSI Caseload in the Context of the Broader Policy and Demographic Landscape*. Cambridge, MA: National Bureau of Economic Research.

Almond, Douglas, and Janet Currie. 2011. "Killing Me Softly: The Fetal Origins Hypothesis." *Journal of Economic Perspectives* 25 (3): 153–172.

Anderson, Mary A., Gina Livermore, AnnaMaria McCutcheon, Todd Honeycutt, Karen Katz, Joseph Mastrianni, and Jacqueline Kauff. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): ASPIRE Process Analysis Report.* Washington, DC: Mathematica Policy Research.

Anderson, Catherine, Ellie Hartman, and D. J. Ralston. 2021. "The Family Empowerment Model: Improving Employment for Youth Receiving Supplemental Security Income." Washington, DC: US Department of Labor, Office of Disability Employment Policy.

Anderson, Catherine A., Amanda Schlegelmilch, and Ellie Hartman. 2019. "Wisconsin PROMISE Cost-Benefit Analysis and Sustainability Framework." *Journal of Vocational Rehabilitation* 51 (2): 253–261.

Anderson, Michael, Yonatan Ben-Shalom, David Stapleton, and David Wittenburg. 2020. *The RETAIN Demonstration: Practical Implications of State Variation in SSDI Entry.* Report for Social Security Administration. Washington, DC: Mathematica Policy Research.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–455.

Arnold Ventures. 2020, December 15. "National RCT of 'Year Up' Program Finds Major, Five-Year Earnings Gains for Low-Income, Minority Young Adults." Straight Talk on Evidence. https://www.straighttalkonevidence.org/2020/12/15/national-rct-of-year-up-program-finds-major-five-year-earnings-gains-for-low-income-minority-young-adults/.

Ashenfelter, O., and M. W. Plant. 1990. "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs." *Journal of Labor Economics* 8 (1): S396-S415.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Autor, David H., and Mark G. Duggan. 2000. "The Rise in Disability Rolls and the Decline in Unemployment." *Quarterly Journal of Economics* 118 (1): 157–205.

Autor, David H., and Mark G. Duggan. 2006. "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding." *Journal of Economic Perspectives* 20 (3): 71–96.

Autor, David, H., and Mark G. Duggan. 2007. "Distinguishing Income from Substitution Effects in Disability Insurance." *American Economic Review* 97 (2): 119–124.

Autor, David H., and Mark Duggan. 2010. *Supporting Work: A Proposal for Modernizing the US Disability Insurance System*. Washington, DC: Center for American Progress and the Hamilton Project.

Autor, David H., Mark G. Duggan, Kyle Greenberg, and David S Lyle. 2016. "The Impact of Disability Benefits on Labor Supply: Evidence from the VA's Disability Compensation Program." *American Economic Journal: Applied Economics* 8 (3): 31–68.

Autor, David H., Nicole Maestas, Kathleen J. Mullen, and Alexander Strand. 2015. *Does Delay Cause Decay? The Effect of Administrative Decision Time on the Labor Force Participation and Earnings of Disability Applicants*. Cambridge, MA: National Bureau of Economic Research.

Autor, David, Nicole Maestas, and Richard Woodberry. 2020. "Disability Policy, Program Enrollment, Work, and Well-Being among People with Disabilities." *Social Security Bulletin* 80 (1): 57.

Bailey, Michelle Stegman, Debra Goetz Engler, and Jeffrey Hemmeter. 2016. "Homeless with Schizophrenia Presumptive Disability Pilot Evaluation." *Social Security Bulletin* 76 (1): 1–25.

Bailey, Michelle Stegman, and Jeffrey Hemmeter. 2015. "Characteristics of Noninstitutionalized DI and SSI Program Participants, 2013 Update." *Social Security Administration Research and Statistics Notes.* No. 2015-02. Social Security Administration. https://www.ssa.gov/policy/docs/rsnotes/rsn2015-02.html.

Bailey, Michelle Stegman, and Robert R. Weathers II. 2014. "The Accelerated Benefits Demonstration: Impacts on Employment of Disability Insurance Beneficiaries." *American Economic Review: Papers & Proceedings* 104 (5): 336–341.

Baller, Julia B., Crystal R. Blyler, Svetlana Bronnikov, Haiyi Xie, Gary R. Bond, Kai Filion, and Thomas Hale. 2020. "Long-Term Follow-up of a Randomized Trial of Supported Employment for SSDI Beneficiaries with Mental Illness." *Psychiatric Services* 71 (3): 243–249.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies." *Journal of Economic Perspectives* 31 (4): 73–102.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *The Annual Review of Economics* 1 (1):151–178.

Barden, Bret. 2013. *Assessing and Serving TANF Recipients with Disabilities*. OPRE Report 2013–56. Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Barnow, Burt S. 1976. "The Use of Proxy Variables When One or Two Independent Variables Are Measured with Error." *American Statistician* 30 (3): 119–121.

Barnow, Burt S., and David Greenberg. 2015. "Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence from Previous Social Experiments." *Evaluation Review* 39 (2): 179–228.

Barnow, Burt S., and David Greenberg. 2019. "Special Issue Editors' Essay." *Evaluation Review* 43 (5): 231–265.

Barnow, Burt S., and David H. Greenberg. 2020. "Conducting Evaluations Using Multiple Trials." *American Evaluation Journal* 41 (4): 529–546.

Bell, Stephen H., and Laura R. Peck. 2016a. "On the Feasibility of Extending Social Experiments to Wider Applications." *Journal of MultiDisciplinary Evaluation* 12 (27): 93–112.

Bell, Stephen H., and Laura R. Peck. 2016b. "On the 'How' of Social Experiments: Experimental Designs for Getting Inside the Black Box." In *Social Experiments in Practice: The What, Why, When, Where, and How of Experimental Design & Analysis*, edited by Laura R. Peck, 97–109. Hoboken, NJ: Jossey-Bass.

Ben-Shalom, Yonatan, Steve Bruns, Kara Contreary, and David Stapleton. 2017. *Stay-at-Work/Return-to-Work: Key Facts, Critical Information Gaps, and Current Practices and Proposal*s. Washington, DC: Mathematica Policy Research.

Ben-Shalom, Yonatan, Jennifer Christian, and David Stapleton. 2018. "Reducing Job Loss among Workers with New Health Problems." In *Investing in America's Workforce: Improving Outcomes for Workers and Employers*, edited by Carl E. Van Horn, 267–288. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Benítiz-Silva, Hugo, Moshe Buchinsky, and John Rust. 2010. "Induced Entry Effects of a $1 for $2 Offset in SSDI Benefits." Mimeo. https://editorialexpress.com/jrust/crest_lectures/induced_entry.pdf.

Berkowitz, E. D. 2013. *The Other Welfare: Supplemental Security Income and US Social Policy.* Ithaca, IL: Cornell University Press.

Berkowitz, Edward D. 2020. *Making Social Welfare Policy in America: Three Case Studies since 1950.* Chicago: University of Chicago Press.

Berkowitz, Edward D., and Larry DeWitt. 2013. *The Other Welfare: Supplemental Security Income and US Social Policy.* Ithaca, NY: Cornell University Press.

Bernanke, Ben. 2012. "The Federal Reserve and the Financial Crisis: Origins and Mission of the Federal Reserve, Lecture 1." Lecture presented at The George Washington University School of Business, Washington, DC, March 20. https://www.federalreserve.gov/mediacenter/files/chairman-bernanke-lecture1-20120320.pdf.

Bezanson, Birdie J. 2004. "The Application of Solution-Focused Work in Employment Counseling." *Journal of Employment Counseling* 41 (4): 183–191.

Biden, J. 2021. *Executive Order on Advancing Racial Equity and Support for Underserved Communities through the Federal Government.* EO 13985. Washington, DC: The White House.

Bitler, Marianne, P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96 (4): 988–1012.

Black, Dan, Kermit Daniel, and Seth Sanders. 2002. "The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust." *American Economic Review* 92 (1): 27–50.

Bloom, Howard S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8 (2): 225–246.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Power of Experimental Designs." *Evaluation Review* 19 (5): 547–566.

Bloom, Howard S. 2009. *Modern Regression Discontinuity Analysis.* New York: MDRC.

Bloom, Howard S., Carolyn J. Hill, and James A. Riccio. 2003. "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments." *Journal of Policy Analysis and Management* 22 (4): 551–575.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources* 32 (3): 549–576.

BLS (Bureau of Labor Statistics), US Department of Labor. 2019. "Characteristics of Unemployment Insurance Applicants and Benefit Recipients – 2018." News Release USDL-19-1692. https://www.bls.gov/news.release/pdf/uisup.pdf.

BLS (Bureau of Labor Statistics), US Department of Labor. 2020a. "Employee Access to Disability Insurance Plans." *The Economics Daily.* https://www.bls.gov/opub/ted/2018/employee-access-to-disability-insurance-plans.htm.

BLS (Bureau of Labor Statistics), US Department of Labor. 2020b. "Employer Reported Workplace Injuries and Illnesses – 2019." News Release USDL-20-2030. https://www.bls.gov/news.release/archives/osh_11042020.pdf.

Blustein, Jan. 2005. "Toward a More Public Discussion of the Ethics of Federal Social Program Evaluation." *Journal of Policy Analysis and Management* 24 (4): 824–846.

Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. 2014. *The 2014 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds*. https://www.ssa.gov/OACT/TR/2014/.

Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. 2019. *The 2019 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds*. Washington, DC: Author. https://www.ssa.gov/oact/tr/2019/tr2019.pdf.

Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. 2021. *The 2021 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds.* Social Security Administration. https://www.ssa.gov/OACT/TR/2021/tr2021.pdf.

Boat, Thomas F., Stephen L. Buka, and James M. Perrin. 2015. "Children with Mental Disorders Who Receive Disability Benefits: A Report from the IOM." *Journal of the American Medical Association* 314 (19): 2019–2020.

Bond, Gary R. 1998. "Principles of the Individual Placement and Support Model: Empirical Support." *Psychiatric Rehabilitation Journal* 22 (1): 11–23.

Bond, G. R., D. R. Becker, and R. E. Drake. 2011. "Measurement of Fidelity of Implementation of Evidence-Based Practices: Case Example of the IPS Fidelity Scale." *Clinical Psychology: Science and Practice* 18: 126–141.

Bond, Gary R., Robert E. Drake, and Deborah R. Becker. 2008. "An Updated on Randomized Control Trials of Evidence-Based Supported Employment." *Psychiatric Rehabilitation. Journal* 31 (4): 280–290.

Bond, Gary R., Robert E. Drake, and Deborah R. Becker. 2012. "Generalizability of the Individual Placement and Support (IPS) Model of Supported Employment Outside the US." *World Psychiatry* 11 (1): 32–-39.

Bond, Gary R., Robert E. Drake, Kim T. Mueser, and Eric Latimer. 2001. "Assertive Community Treatment for People with Severe Mental Illness." *Disease Management and Health Outcomes* 9 (3): 141–159.

Bond, Gary R., Robert E. Drake, and Jacqueline A. Pogue. 2019. "Expanding Individual Placement and Support to Populations 441hat Conditions and Disorders Other Than Serious Mental Illness." *Psychiatric Services* 70 (6): 488–498.

Bound, John. 1989. "The Health and Earnings of Rejected Disability Insurance Applicants." *American Economic Review* 79 (3): 482–503.

Bound, John. 1991. "The Health and Earnings of Disability Insurance Applicants: Reply." *American Economic Review* 81 (5): 1427–1434.

Bound, John, and Richard V. Burkhauser. 1999. "Economic Analysis of Transfer Programs Targeted on People with Disabilities." In *Handbook of Labor Economics*, vol. 3, edited by Orley Ashenfelter and David Card, 3417–3528. Amsterdam, The Netherlands: Elsevier.

Bound, John, Richard V. Burkhauser, and Austin Nichols. 2003. "Tracking the Household Income of SSDI and SSI Applicants." *Research in Labor Economics* 22: 113–158.

Bound, John, Julie Berry Cullen, Austin Nichols, and Lucie Schmidt. 2004. "The Welfare Implications of Increasing Disability Insurance Benefit Generosity." *Journal of Public Economics* 88 (12): 2487–2514.

Bound, John, Stephan Lindner, and Tim Waidmann. 2014. "Reconciling Findings on the Employment Effect of Disability Insurance." *IZA Journal of Labor Policy* 3 (1): 1–23.

Boyer, Sara L., and Gary R. Bond. 1999. "Does Assertive Community Treatment Reduce Burnout? A Comparison with Traditional Case Management." *Mental Health Services Research* 1 (1): 31–45.

Braitman, Alex, Peggy Counts, Richard Davenport, Barbara Zurlinden, Mark Rogers, Joe Clauss, Arun Kulkarni, Jerry Kymla, and Laura Montgomery. 1995. "Comparison of Barriers to Employment for Unemployed and Employed Clients in a Case Management Program: An Exploratory Study." *Psychiatric Rehabilitation Journal* 19 (1): 3–8.

Brock, Thomas, Michael J. Weiss, and Howard S. Bloom. 2013. *A Conceptual Framework for Studying the Sources of Variation in Program Effects.* New York: MDRC.

Brownson, Ross C., Amy A. Eyler, Jenine K. Harris, Justin B. Moore, and Rachel G. Tabak. 2018. "Getting the Word Out: New Approaches for Disseminating Public Health Science." *Journal of Public Health Management and Practice* 24 (2): 102–111.

Bruyere, Susanne M., Thomas P. Golden, and Ilene Zeitzer. 2007. "Evaluation and Future Prospect of U.S. Return to Work Policies for Social Security Beneficiaries." *Disability and Employment* 59: 53–90.

Burkhauser, Richard V., and Mary C. Daly. 2011. *The Declining Work and Welfare of People with Disabilities: What Went Wrong and a Strategy for Change.* Washington, DC: American Enterprise Institute Press.

Burkhauser, Richard V., Mary C. Daly, Duncan McVicar, and Roger Wilkins. 2014. "Disability Benefit Growth and Disability Reform in the US: Lessons from other OECD Nations." *IZA Journal of Labor Policy* 3 (4): 1–30.

Burstein, Nancy R., Cheryl A. Roberts, and Michelle L. Wood. 1999. *Recruiting SSA's Disability Beneficiaries for Return-to-Work: Results of the Project NetWork Demonstration: Final Report.* Bethesda, MD: Abt Associates.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research." *The Journal of Economic Perspectives* 9 (2): 63–84.

Burtless, Gary, and David Greenberg. 1982. "Inferences Concerning Labor Supply Behavior Based on Limited Duration Experiments." *The American Economic Review* 72 (3): 488–497.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22 (1): 31–72.

Camacho, Christa Bucks, and Jeffrey Hemmeter. 2013. "Linking Youth Transition Support Services: Results from Two Demonstration Projects." *Social Security Bulletin* 73 (1). https://www.ssa.gov/policy/docs/ssb/v73n1/v73n1p59.html.

Campbell, Frances A., Elizabeth P. Pungello, Shari Miller-Johnson, Margaret Burchinal, and Craig T. Ramey. 2001. "The Development of Cognitive and Academic Abilities: Growth Curves from an Early Childhood Educational Experiment." *Developmental Psychology* 37 (2): 231–242.

Card, David, Jochen Kluve, and Andrea Weber. 2010. "Active Labour Market Policy Evaluations: A Meta-Analysis." *The Economic Journal* 120 (548): F452–F477.

Carter, Erik W., Diane Austin, and Audrey A. Trainor. 2012. "Predictors of Postschool Employment Outcomes for Young Adults with Severe Disabilities." *Journal of Disability Policy Studies* 23 (1): 50–63.

CBPP (Center on Budget and Policy Priorities). 2021. *Supplemental Security Income.* Policy Basics. Washington, DC: Author. https://www.cbpp.org/sites/default/files/atoms/files/PolicyBasics_SocSec-IntroToSSI.pdf.

CEA (Council of Economic Advisers). 2016. *Economic Report of the President, Transmitted to the Congress February 2016 Together with the Annual Report of the Council of Economic Advisors.* Washington DC: Government Printing Office.

CEP (Commission on Evidence-Based Policymaking). 2017. *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking.* Washington, DC: Author. https://bipartisanpolicy.org/wp-content/uploads/2019/03/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Comission-on-Evidence-based-Policymaking.pdf.

Chambless, Cathy, George Julnes, Sara McCormick, and Anne Brown-Reither. 2009. *Utah SSDI $1 for $2 Benefit Offset Pilot Demonstration Final Report.* Salt Lake City, UT: State of Utah.

Chambless, Catherine E., George Julnes, Sara T. McCormick, and Anne Reither. 2011. "Supporting Work Effort of SSDI Beneficiaries: Implementation of Benefit Offset Pilot Demonstration." *Journal of Disability Policy Studies* 22 (3): 179–188.

Charles, Kerwin Kofi, Yiming Li, and Melvin Stephens, Jr. 2018. "Disability Benefit Take-Up and Local Labor-Market Conditions." *Review of Economics and Statistics* 100 (3): 416–423.

Chetty, Raj. 2006. "A General Formula for the Optimal Level of Social Insurance." *Journal of Public Economics* 90 (10): 1879–1901.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. 2017. "The Fading American Dream: Trends in Absolute Income Mobility since 1940." *Science* 356 (6336): 398–406.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106 (4): 855–902.

Chow, Shein-Chung, and Mark Chang. 2012. *Adaptive Design Methods in Clinical Trials.* 2nd ed. Boca Raton, FL: CRC Press.

Christian, Jennifer, Thomas Wickizer, and A. Kim Burton. 2016. "A Community-Focused Health & Work Service (HWS)." In *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program*, edited by Committee for a Responsible Federal Budget, The McCrery-Pomeroy SSDI Solutions Initiative, Ch. 4. Offprint. https://www.crfb.org/sites/default/files/christianwickizerburton.pdf.

Committee for a Responsible Federal Budget, The McCrery-Pomeroy SSDI Solutions Initiative. 2016. *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program.* West Conshohocken, PA: Infinity Publishing.

Claes, Rita, and S. Antonio Ruiz-Quintanilla. 1998. "Influences of Early Career Experiences, Occupational Group, and National Culture on Proactive Career Behavior." *Journal of Vocational Behavior* 52 (3): 357–378.

Cloutier, Heidi, Joanne Malloy, David Hagner, and Patricia Cotton. 2006. "Choice and Control over Resources: New Hampshire's Individual Career Account Demonstration Projects." *Journal of Rehabilitation* 72 (2): 4–11.

Coldwell, Craig M., and William S. Bender. 2007. "The Effectiveness of Assertive Community Treatment for Homeless Populations with Severe Mental Illness: A Meta-Analysis." *American Journal of Psychiatry* 164 (3): 393–399.

Committee for the Prize in Economic Sciences in Memory of Alfred Nobel. 2019. *Understanding Development and Poverty Alleviation.* Stockholm, Sweden: The Royal Swedish Academy of Sciences.

Congressional Budget Office. 2012. *Policy Options for the Social Security Disability Insurance Program.* Washington, DC: Congress of the United States, Congressional Budget Office.

Cook, Thomas D. 2018. "Twenty-Six Assumptions That Have to Be Met If Single Random Assignment Experiments Are to Warrant 'Gold Standard' Status: A Commentary on Deaton and Cartwright." *Social Science & Medicine* 210: 37–40.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27 (4): 724–750.

Cook, J., S. Shore, J. Burke-Miller, J. Jonikas, M. Hamilton, B. Ruckdeschel, et al. 2019. "Efficacy of Mental Health Self-Directed Care Financing in Improving Outcomes and Controlling Service Costs for Adults with Serious Mental Illness." *Psychiatric Services* 70 (3): 191–201.

Costa, Jackson. 2017. "The Decline in Earnings Prior to Application for Disability Insurance Benefits. *Social Security Bulletin* 77(1). https://www.ssa.gov/policy/docs/ssb/v77n1/v77n1p1.html.

Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Polices Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 1238 (2): 531–580.

Cronbach, Lee J., Sueann Robinson Ambron, Sanford M. Dornbusch, Robert C. Hornik, D. C. Phillips, Decker F. Walker, and Stephen S. Winer. 1980. *Toward Reform of Program Evaluation.* San Francisco: Jossey-Bass.

Cunha, Flavio, and James J. Heckman. 2007. "The Evolution of Inequality, Heterogeneity, and Uncertainty in Labor Earnings in the US Economy." NBER Paper No. 13526. Cambridge, MA: National Bureau of Economic Research.

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43 (4): 738–782.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." NBER Paper No. 11331. Cambridge, MA: National Bureau of Economic Research.

Davies, Paul S., Kalman Rupp, and David Wittenburg. 2009. "A Life-Cycle Perspective on the Transition to Adulthood among Children Receiving Supplemental Security Income Payments." *Journal of Vocational Rehabilitation* 30 (3): 133–151.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

Decker, Paul T., and Craig V. Thornton. 1995. "The Long-Term Effects of Transitional Employment Services." *Social Security Bulletin* 58 (4): 71–81.

Delin, Barry S., Ellie C. Hartman, and Christopher W. Sell. 2012. "The Impact of Work Outcomes: Evidence from Two Return-to-Work Demonstrations." *Journal of Vocational Rehabilitation* 36 (2): 97–107.

Delin, Barry S., Ellie C. Hartman, Christopher W. Sell, and Anne E. Brown-Reither. 2010. *Testing a SSDI Benefit Offset: An Evaluation of the Wisconsin SSDI Employment Pilot.* Menomonie, WI: University of Wisconsin-Stout.

Denne, Jacob, George Kettner, and Yonatan Ben-Shalom. 2015. *Return to Work in the Health Care Sector: Promising Practices and Success Stories.* Report for US Department of Labor, Office of Disability Employment Policy. Washington, DC: Mathematica Policy Research.

Derr, Michelle, Denise Hoffman, Jillian Berk, Ann Person, David Stapleton, Sarah Croake, Christopher Jones, and Jonathan McCay. 2015. *BOND Implementation and Evaluation: Process Study Report.* Washington, DC: Mathematica Policy Research.

Deshpande, Manasi. 2016a. "Does Welfare Inhibit Success? I Long-Term Effects of Removing Low-Income Youth from the Disability Rolls." *American Economic Review* 106 (11): 3300–3330.

Deshpande, Manasi. 2016b. "The Effect of Disability Payments on Household Earnings and Income: Evidence from the SSI Children's Program." *Review of Economics and Statistics* 98 (4): 638–654.

Deshpande, Manasi. 2020. "How Disability Benefits in Early Life Affect Long-Term Outcomes." Center Paper NB20-05. Cambridge, MA: National Bureau of Economic Research.

Deshpande, Manasi, and Rebecca Dizon-Ross. 2020. *Improving the Outcomes of Disabled Youth through Information.* Cambridge, MA: National Bureau of Economic Research. https://grantome.com/grant/NIH/R21-HD091472-02.

DiClemente, Carlo C., James O. Prochaska, Scott K. Fairhurst, Wayne F. Velicer, Mary M. Velasquez, and Joseph S. Rossi. 1991. "The Process of Smoking Cessation: An Analysis of Precontemplation, Contemplation, and Preparation Stages of Change." *Journal of Consulting and Clinical Psychology* 59 (2): 295–304.

DiNardo, John, Jordan Matsudaira, Justin McCrary, and Lisa Sanbonmatsu. 2021. "A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example." *Journal of Labor Economics* 39 (S2): S507–S541.

Dixon, Lisa. 2000. "Assertive Community Treatment: Twenty-Five Years of Gold." *Psychiatric Services* 51 (6): 759–765.

Doemeland, Doerte, and James Trevino. 2014. "Which World Bank Reports Are Widely Read?" World Bank Policy Research Working Paper No. 6851. Washington, DC: The World Bank. http://documents1.worldbank.org/curated/en/387501468322733597/pdf/WPS6851.pdf.

DOL (US Department of Labor). 2015 [updated 2019]. *CLEAR Causal Evidence Guidelines, Version 2.1*. Washington, DC: US Department of Labor, Clearinghouse for Labor Evaluation and Research. https://clear.dol.gov/reference-documents/causal-evidence-guidelines-version-21.

DOL (US Department of Labor). n.d. "Employment First Presents 10 Critical Areas for Improving Competitive Integrated Employment Based on the WIOA Advisory Committee Report." Accessed December 10, 2020. https://www.dol.gov/sites/dolgov/files/odep/topics/employmentfirst/ef-presents-10-critical-areas-for-improving-cie-based-on-the-wioa-advisory-committee-report-full.pdf.

DOL (US Department of Labor). n.d. "RETAIN Initiative." Accessed September 24, 2021. https://www.dol.gov/agencies/odep/initiatives/saw-rtw/retain.

DOL (US Department of Labor). n.d. "WIOA Title I and III Annual Report Data: Program Year 2019." Workforce Performance Results, Employment and Training Administration. Accessed May 12, 2021. https://www.dol.gov/agencies/eta/performance/results.

DOL (US Department of Labor), ODEP (Office of Disability Employment Policy). 2018. "Notice of Availability of Funds and Funding Opportunity Announcement for: Retaining Employment and Talent after Injury/Illness Network Demonstration Projects." Issued May 24, 2018. https://www.dol.gov/sites/dolgov/files/odep/topics/saw-rtw/docs/foa-odep-18-01-published-on-grants.gov.pdf.

Dong, Nianbo, and Rebecca Maynard. 2013. "PowerUp! A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies." *Journal of Research on Educational Effectiveness* 6 (1): 24–67.

Duggan, Mark, and Scott A. Imberman. 2009. "Why Are the Disability Rolls Skyrocketing? The Contribution of Population Characteristics, Economic Conditions, and Program Generosity." In *Health at Older Ages*, edited by David M. Cutler and David A. Wise, 337–380. Chicago: University of Chicago Press.

Duggan, Mark G., and Melissa S. Kearney. 2007. "The Impact of Child SSI Enrollment on Household Outcomes." *Journal of Policy Analysis and Management* 26 (4): 861–885.

Duggan, Mark, Melissa S. Kearney, and Stephanie Rennane. 2015. "The Supplemental Income (SSI) Program." NBER Working Paper No. 21209. Cambridge, MA: National Bureau of Economic Research.

Duggan, Mark, Melissa S. Kearney, and Stephanie Rennane. 2016. "The Supplemental Security Income Program." In *Economics of Means-Tested Transfer Programs in the United Stated,* Vol. 2, edited by Robert A. Moffitt, 1–58. Chicago: University of Chicago Press.

Durlak, Joseph A., and Emily P. DuPre. 2008. "Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation." *American Journal of Community Psychology* 41 (3): 327-350.

Eeckhoudt, Louis, and Miles Kimball. 1992. "Background Risk, Prudence, and the Demand for Insurance." In *Contributions to Insurance Economics*, edited by Georges Dionne, 23–54. Boston: Kluwer Academic Publishers.

Eichengreen, Barry. 1996. *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939*. New York: Oxford University Press.

Ekman, Lisa D. 2016. "Discussion of Early Intervention Proposals." In *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program*, edited by Committee for a Responsible Federal Budget, The McCrery-Pomeroy SSDI Solutions Initiative, Ch. 3. Offprint. https://www.crfb.org/sites/default/files/stapletonbenshalommann.pdf.

Ellenhorn, Ross. 2005. "Parasuicidality and Patient Careerism: Treatment Recidivism and the Dialectics of Failure." *American Journal of Orthopsychiatry* 75 (2): 288–303.

Ellison, Marsha Langer, E. Sally Rogers, Ken Sciarappa, Mikal Cohen, and Rick Forbess. 1995. "Characteristics of Mental Health Case Management: Results of a National Survey." *The Journal of Mental Health Administration* 22 (2): 101–112.

Epstein, Diana, and Jacob Alex Klerman. 2012. "When Is a Program Ready for Rigorous Impact Evaluation? The Role of a Falsifiable Logic Model." *Evaluation Review* 36 (5): 375–401.

Epstein, Z., M. Wood, M. Grosz, S. Prenovitz, and A. Nichols. 2020. *Synthesis of Stay-at-Work/Return-to-Work (SAW/RTW) Programs, Models, Efforts, and Definitions.* Cambridge, MA: Abt Associates.

Farrell, Mary, Peter Baird, Bret Barden, Mike Fishman, and Rachel Pardoe. 2013. *The TANF/SSI Disability Transition Project: Innovative Strategies for Serving TANF Recipients with Disabilities*. OPRE Report 2013-51. Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Farrell, Mary, and Johanna Walter. 2013. *The Intersection of Welfare and Disability: Early Findings from the TANF/SSI Disability Transition Project*. OPRE Report 2013-06. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Feely, Megan, Kristen D. Seay, Paul Lanier, Wendy Auslander, and Patricia L. Kohl. 2018. "Measuring Fidelity in Research Studies: A Field Guide to Developing a Comprehensive Fidelity Measurement System." *Child and Adolescent Social Work Journal* 35 (2): 139–152.

Fein, David, Samuel Dastrup, and Kimberly Burnett. 2021. *Still Bridging the Opportunity Divide for Low-Income Youth: Year Up's Longer-Term Impacts*. OPRE Report 2021-56. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services. https://www.acf.hhs.gov/sites/default/files/documents/opre/year-up-report-april-2021.pdf.

Finkelstein, Amy, and Nathaniel Hendren. 2020. "Welfare Analysis Meets Causal Inference." *Journal of Economic Perspectives* 34 (4): 146–67. https://doi.org/10.1257/jep.34.4.146

Finkelstein, Amy, Sarah Taubman, Heidi Allen, Jonathan Gruber, Joseph P. Newhouse, Bill Wright, Kate Baicker, and Oregon Health Study Group. 2010. "*The Short-Run Impact of Extending Public Health Insurance to Low Income Adults: Evidence from the First Year of the Oregon Medicaid Experiment. Analysis Plan.* https://www.nber.org/sites/default/files/2020-02/analysis-plan-one-year-2010-12-01.pdf.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *The Quarterly Journal of Economics* 127 (3): 1057–1106.

Foster L., R. Brown, P. Phillips, J. Schore, and B. L. Carlson. 2003. "Improving the Quality of Medicaid Personal Assistance through Consumer Direction." *Health Affairs* 22 (Suppl 1). https://doi.org/10.1377/hlthaff.w3.162.

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. 2011. "Subgroup Identification from Randomized Clinical Trial Data." *Statistics in Medicine* 30 (24): 2867–2880. https://doi.org/10.1002/sim.4322.

Fraker, Thomas M., Peter Baird, Alison Black, Arif Mamun, Michelle Manno, John Martinez, Anu Rangarajan, and Debbie Reed. 2011. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on Colorado Youth WIN*. Report for Social Security Administration, Office of Program Development and Research. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Peter Baird, Arif Mamun, John Martinez, Debbie Reed, and Allison Thompkins. 2012. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on the Career Transition Program.* Center for Studying Disability Policy. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Alison Black, Joseph Broadus, Arif Mamun, Michelle Manno, John Martinez, Reanin McRoberts, Anu Rangarajan, and Debbie Read. 2011. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on the City University of New York's Project.* Center for Studying Disability Policy. Washington, DC: Mathematica Policy Research.

Fraker, Thomas M., Alison Black, Arif Mamun, Michelle Manno, John Martinez, Bonnie O'Day, Meghan O'Toole, Anu Rangarajan, and Debbie Reed. 2011" *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on Transition WORK"*. Report for Social Security Administration, Office of Program Development and Research. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Alison Black, Arif Mamun, John Martinez, Bonnie O'Day, Meghan O'Toole, Anu Rangarajan, and Debbie Read. 2011. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on the Transition Works Project.* Center for Studying Disability Policy. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Erik Carter, Todd Honeycutt, Jacqueline Kauff, Gina Livermore, and Arif Mamun. 2014. *Promoting Readiness of Minors in SSI (PROMISE) Evaluation Design Report*. Washington, DC: Mathematica Policy Research.

Fraker, Thomas M., Joyanne Cobb, Jeffrey Hemmeter, Richard G. Luecking, and Arif Mamun. 2018. "Three-Year Effects of the Youth Transition Demonstration Projects." *Social Security Bulletin* 78 (3): 19–41.

Fraker, Thomas, Todd Honeycutt, Arif Mamun, Michelle Manno, John Martinez, Bonnie O'Day, Debbie Reed, and Allison Thompkins. 2012. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on the Broadened Horizons, Brighter Futures.* Center for Studying Disability Policy. Washington, DC: Mathematica Policy Research.

Fraker, Thomas M., Richard G. Luecking, Arif A. Mamun, John M. Martinez, Deborah S. Reed, and David C. Wittenburg. 2016. "An Analysis of 1-Year Impacts of Youth Transition Demonstration Projects." *Career Development and Transition for Exceptional Individuals* 39 (1): 34–46.

Fraker, Thomas, Arif Mamun, Todd Honeycutt, Allison Thompkins, and Erin J. Valentine. 2014. *Final Report on the Youth Transition Demonstration*. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Arif Mamun, Michelle Manno, John Martinez, Debbie Reed, Allison Thompkins, and David Wittenburg. 2012. *The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on the West Virginia Youth Works Project.* Center for Studying Disability Policy. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, Arif Mamun, and Lori Timmins. 2015. *Three-Year Impacts of Services and Work Incentives on Youth with Disabilities*. Washington, DC: Mathematica Policy Research.

Fraker, Thomas, and Anu Rangarajan. 2009. "The Social Security Administration's Youth Transition Demonstration Projects." *Journal of Vocational Rehabilitation* 30 (3): 223–240.

Francesconi, Marco, and James J. Heckman. 2016. "Child Development and Parental Investment: Introduction." *The Economic Journal* 126 (596): F1–F27. https://doi.org/10.1111/ecoj.12388.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.

Franklin, Gary M., Thomas M. Wickizer, Norma B. Coe, and Deborah Fulton-Kehoe. 2015. "Workers' Compensation: Poor Quality Health Care and the Growing Disability Problem in the United States." *American Journal of Industrial Medicine* 58 (3): 245–251.

Freburger, Janet K., George M. Holmes, Robert P. Agans, Anne M. Jackman, Jane D. Darter, Andrea S. Wallace, Liana D. Castel, William D. Kalsbeek, and Timothy S. Carey. 2009. "The Rising Prevalence of Chronic Low Back Pain." *Archives of Internal Medicine* 169 (3): 251–258.

Freedman, Lily, Sam Elkin, and Megan Millenky. 2019. "Breaking Barriers: Implementing Individual Placement and Support in a Workforce Setting." New York: MDRC.

French, Eric, and Jae Song. 2014. "The Effect of Disability Insurance Receipt on Labor Supply." *American Economic Journal: Economic Policy* 6 (2): 291–337.

Frey, William D., Robert E. Drake, Gary R. Bond, Alexander L. Miller, Howard H. Goldman, David S. Salkever, Steven Holsenbeck, Mustafa Karakus, Roline Milfort, Jarnee Riley, Cheryl Reidy, Julie Bollmer, and Megan Collins. 2011. *Mental Health Treatment Study: Final Report*. Rockville, MD: Westat.

Fukui, Sadaaki, Rick Goscha, Charles A. Rapp, Ally Mabry, Paul Liddy, and Doug Marty. 2012. "Strengths Model Case Management Fidelity Scores and Client Outcomes." *Psychiatric Services* 63 (7): 708–710.

GAO (US Government Accountability Office). 2002. *Program Evaluation: Strategies for Assessing How Information Dissemination Contributes to Agency Goals*. Report No. GAO-02-923. Washington, DC: Author.

GAO (US Government Accountability Office). 2004. *Social Security Disability: Improved Processes for Planning and Conducting Demonstrations May Help SSA More Effectively Use Its Demonstration Authority.* Report No. GAO-05-19. Washington, DC: Author.

GAO (US Government Accountability Office). 2005. *Federal Disability Assistance, Wide Array of Programs Needs to Be Examined in Light of 21st Century Challenges*. Report No. GAO-05-626. Washington, DC: Author.

GAO (US Government Accountability Office). 2008. *Social Security Disability: Management Controls Needed to Strengthen Demonstration Projects*. Report No. GAO-08-1053. Washington, DC: Author.

GAO (US Government Accountability Office). 2010. *Highlights of a Forum: Actions That Could Increase Work Participation for Adults with Disabilities*. Report No. GAO-10-812SP. Washington, DC: Author.

GAO (US Government Accountability Office). 2012a. *Designing Evaluations: 2012 Revision*. Report No. GAO-12-208G. Washington, DC: Author.

GAO (US Government Accountability Office). 2012b. *Employment for People with Disabilities: Little Is Known about the Effectiveness of Fragmented and Overlapping Programs*. Report No. GAO-12-677. Washington, DC: Author.

GAO (US Government Accountability Office). 2012c. *Supplemental Security Income: Better Management Oversight Needed for Children's Benefits*. Report No. GAO-12-498SP. Washington, DC: Author.

GAO (US Government Accountability Office). 2017. *Supplemental Security Income: SSA Could Strengthen Its Efforts to Encourage Employment for Transition-Age Youth*. Report No. GAO-17-485. Washington, DC: Author.

GAO (US Government Accountability Office). 2018. *Medicaid Demonstrations: Evaluations Yielded Limited Results, Underscoring Need for Changes to Federal Policies and Procedures*. Report No. GAO-18-220. Washington, DC: Author.

GAO (US Government Accountability Office). 2019. *Medicaid Demonstrations: Approvals of Major Changes Need Increased Transparency*. Report No. GAO-19-315. Washington, DC: Author.

Gardiner, Karen N., and Randall Juras. 2019. *Pathways for Advancing Careers and Education: Cross-Program Implementation and Impact Study Findings*. OPRE Report 2019-32. Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Gary, K. W., A. Sima, P. Wehman, and K. R. Johnson. 2019. "Transitioning Racial/Ethnic Minorities with Intellectual and Developmental Disabilities: Influence of Socioeconomic Status on Related Services." *Career Development and Transition for Exceptional Individuals* 42 (3): 158–167. https://doi.org/10.1177/2165143418778556.

Gelber, Alexander, Timothy J. Moore, and Alexander Strand. 2017. "The Effect of Disability Insurance Payments on Beneficiaries' Earnings." *American Economic Journal: Economic Policy* 9 (3): 229–261.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, DC: The International Bank for Reconstruction and Development, The World Bank.

Geyer, Judy, Daniel Gubits, Stephen Bell, Tyler Morrill, Denise Hoffman, Sarah Croake, Katie Morrison, David Judkins, and David Stapleton. 2018. *BOND Implementation and Evaluation: 2017 Stage 2 Interim Process, Participation, and Impact Report.* Report for the Social Security Administration. Cambridge, MA: Abt Associates.

Gimm, Gilbert, Noelle Denny-Brown, Boyd Gilman, Henry T. Ireys, and Tara Anderson. 2009. *Interim Report on the National Evaluation of the Demonstration to Maintain Independence and Employment.* Washington, DC: Mathematica Policy Research.

Gingerich, Jade Ann, and Kelli Crane. 2021. *Transition Linkage Tool: A System Approach to Enhance Post-School Employment Outcomes.* Washington, DC: US Department of Labor, Office of Disability Employment Policy.

Gokhale, Jagadeesh. 2013" *A New Approach to SSDI Reform."* McCrery-Pomeroy SSDI Solutions Initiative Policy Brief. Washington, DC: Committee for a Responsible Federal Budget.

Gokhale, Jagadeesh. 2015. "SSDI Reform: Promoting Return to Work Without Compromising Economic Security." *Wharton Public Policy Initiative* 3 (7): 1–6.

Golden, Thomas P., Susan O'Mara, Connie Ferrell, and James R. Sheldon, Jr. 2000. "A Theoretical Construct for Benefits Planning and Assistance in the Ticket to Work and Work Incentive Improvement Act." *Journal of Vocational Rehabilitation* 14, (3): 147–152. https://content.iospress.com/articles/journal-of-vocational-rehabilitation/jvr00076.

Golden, T. P., S. O'Mara, C. Ferrell, J. Sheldon, and L. Axton Miller. 2005. *Supporting Career Development and Employment: Benefits Planning, Assistance and Outreach (BPA&O) and Protection and Advocacy for Beneficiaries of Social Security (PABSS).* SSA Publication No. 63-003. Social Security Administration. https://hdl.handle.net/1813/89921.

Goss, Steven C. 2013. *Testimony by Chief Actuary from Social Security Administration before the House Committee on Ways and Means, Subcommittee on Social Security*. Washington, DC: Social Security Administration.

Greenberg, David, Genevieve Knight, Stefan Speckesser, and Debra Hevenstone. 2011. "Improving DWP Assessment of the Relative Costs and Benefits of Employment Programmes." Working Paper No. 100. London, England: Department for Work and Pensions.

Greenberg, David, Robert H. Meyer, and Michael Wiseman. 1993. *Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs*. Madison, WI: University of Wisconsin-Madison, Institute for Research on Poverty.

Greenberg, David, Robert H. Meyer, and Michael Wiseman. 1994. "Multi-Site Employment and Training Evaluations: A Tale of Three Studies." *Industrial and Labor Relations Review* 47 (4): 679–691.

GSA (General Services Administration), OES (Office of Evaluation Sciences). 2018. *Increasing SSI Uptake: Letters to Adults 65 and Older Increased SSI Awards by 340%.* Washington, DC: Authors. https://oes.gsa.gov/assets/abstracts/1723-Increasing-SSI-Uptake.pdf.

GSA (General Services Administration), OES (Office of Evaluation Sciences). 2019a. *Communicating Employment Supports to Denied Disability Insurance Applicants.* https://oes.gsa.gov/assets/abstracts/15xx-di.pdf.

GSA (General Services Administration), OES (Office of Evaluation Sciences). 2019b. *Encouraging SSI Recipients to Self-Report Wage Changes.* Washington, DC: Authors. https://oes.gsa.gov/assets/abstracts/XXXX-ssi-wage-reporting-abstract.pdf.

GSA (General Services Administration), OES (Office of Evaluation Sciences). 2019c. "Encouraging SSI Recipients to Self-Report Wage Changes." https://oes.gsa.gov/projects/ssi-wage-reporting/.

Gubits, Daniel, Rachel Cook, Stephen Bell, Michelle Derr, Jillian Berk, Ann Person, David Stapleton, Denise Hoffman, and David Wittenburg. 2013. *BOND Implementation and Evaluation: Stage 2 Early Assessment Report.* Rockville, MD: Abt Associates.

Gubits, Daniel, Judy Geyer, Denise Hoffman, Sarah Croake, Utsav Kattel, David Judkins, Stephen Bell, and David Stapleton. 2017. *BOND Implementation and Evaluation: 2015 Stage 2 Interim Process, Participation, and Impact Report.* Report for Social Security Administration, Office of Program Development & Research. Cambridge, MA: Abt Associates; and Washington, DC: Mathematica Policy Research.

Gubits, Daniel R., Judy Geyer, David Stapleton, David Greenberg, Stephen Bell, Austin Nichols, Michelle Wood, Andrew McGuirk, Denise Hoffman, Meg Carroll, Sarah Croake, Utsav Kattel, David R Mann, and David Judkins. 2018a. *BOND Implementation and Evaluation: Final Evaluation Report,* Vol. 1. Report for the Social Security Administration. Cambridge, MA: Abt Associates; and Washington, DC: Mathematica Policy Research.

Gubits, Daniel R., Judy Geyer, David Stapleton, David Greenberg, Stephen Bell, Austin Nichols, Michelle Wood, Andrew McGuirk, Denise Hoffman, Meg Carroll, Sarah Croake, Utsav Kattel, David Mann, and David Judkins. 2018b. *BOND Implementation and Evaluation: Final Evaluation Report.* Vol. 2, *Technical Appendices*. Report for Social Security Administration. Cambridge, MA: Abt Associates; and Washington, DC: Mathematica Policy Research.

Gubits, Daniel, Sarah Gibson, Michelle Wood, Cara Sierks, and Zachary Epstein. 2019. *Post-Entitlement Earnings Simplification Demonstration Technical Experts Panel Meeting: Final Report*. Rockville, MD: Abt Associates.

Guldi, Melanie, Amelia Hawkins, Jeffrey Hemmeter, and Lucie Schmidt. 2018. "Supplemental Security Income and Child Outcomes: Evidence from Birth Weight Eligibility Cutoffs." NBER Working Paper No. 24913. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w24913.

Hahn, Robert. 2019. "Building upon Foundations for Evidence-Based Policy," *Science* 364 (6440): 534–535.

Hall, Jean P., Catherine Ipsen, Noelle K. Kurth, Sara McCormick, and Catherine Chambless. 2020. "How Family Crises May Limit Engagement of Youth with Disabilities in Services to Support Successful Transitions to Postsecondary Education and Employment." *Children and Youth Services Review* 118: 1–7.

Hammermesh, Daniel S. 2007. "Viewpoint: Replication in Economics." *Canadian Journal of Economics* 40 (3): 715–733.

Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs,* edited by Charles F. Manski and Irwin Garfinkel. Cambridge, MA: Harvard University Press.

Heckman, James J. 2011. "The Economics of Inequality: The Value of Early Childhood Education." *American Educator* 35, no. 1 (Spring): 31–47.

Heckman, James, Lance Lochner, and Ricardo Cossa. 2003. "Learning-by-Doing versus On-the-Job Training: Using Variation Induced by the EITC to Distinguish between Models of Skill Formation." In *Designing Social Inclusion: Tools to Raise Low-End Pay and Employment in Private Enterprise*, edited by Edmund S. Phelps, 74–130. Cambridge, United Kingdom: Cambridge University Press.

Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annual Review of Economics* 6 (1): 689–733.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2): 85–110.

Heckman, James J., and Jeffrey A. Smith. 2004. "The Determinants of Participation in a Social Program: Evidence from a Prototypical Job Training Program." *Journal of Labor Economics* 22 (2): 243–298.

Heckman, James, Jeffrey Smith, and Christopher Taber. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *The Review of Economics and Statistics* 80 (1): 1–14.

Heckman, J. J., and E. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1." *Econometrica*, 73 (3): 669–738.

Hemmeter, Jeffrey. 2014. "Earnings and Disability Program Participation of Youth Transition Demonstration Participants after 24 Months." *Social Security Bulletin* 74 (1). https://www.ssa.gov/policy/docs/ssb/v74n1/v74n1p1.html.

Hemmeter, Jeffrey. 2015. "Supplemental Security Income Program Entry at Age 18 and Entrants' Subsequent Earnings." *Social Security Bulletin* 75 (3): 35–53.

Hemmeter, Jeffrey, and Michelle Stegman Bailey. 2016. "Earnings after DI: Evidence from Full Medical Continuing Disability Reviews." *IZA Journal of Labor Policy* 5 (1): 1–22.

Hemmeter, Jeffrey, and Joyanne Cobb. 2018. *Youth Transition Demonstration: Follow-Up Findings*. Presentation at the Fall Research Conference of the Association for Public Policy Analysis & Management, Washington, DC, November 2018.

Hemmeter, Jeffrey, Mark Donovan, Joyanne Cobb, and Tad Asbury. 2015. "Long Term Earnings and Disability Program Participation Outcomes of the Bridges Transition Program." *Journal of Vocational Rehabilitation* 42 (1): 1–15.

Hemmeter, Jeffrey, Michael Levere, Pragya Singh, and David Wittenburg. 2021. "Changing Stays? Duration of Supplemental Security Income Participation by First-Time Child Awardees and the Role of Continuing Disability Reviews." *Social Security Bulletin* 81 (2): 17–41.

Hemmeter, Jeffrey, David R. Mann, and David C. Wittenburg. 2017. "Supplemental Security Income and the Transition to Adulthood in the United States: State Variations in Outcomes Following the Age-18 Redetermination." *Social Service Review* 91 (1): 106–133.

Hemmeter, Jeffrey, John Phillips, Elana Safran, and Nicholas Wilson. 2020 "Communicating Program Eligibility: A Supplemental Security Income Field Experiment." Office of Evaluation Sciences Working Paper. https://oes.gsa.gov/assets/publications/1723%20-%20Hemmeter%20et%20al%20(2021)%20-%20Communicating%20Program%20Eligibility%20A%20Supplemental%20Security%20Income%20(SSI)%20Field%20Experiment.pdf.

Hemmeter, Jeffrey, and Michelle Stegman. 2015. "Childhood Continuing Disability Reviews and Age-18 Redeterminations for Supplemental Security Income Recipients: Outcomes and Subsequent Program Participation." *Research and Statistics Notes*. No. 2015-03. Social Security Administration. https://www.ssa.gov/policy/docs/rsnotes/rsn2015-03.html

Hendra, R., James A. Riccio, Richard Dorsett, David H. Greenberg, Genevieve Knight, Joan Phillips, Philip K. Robins, Sandra Vegeris, Johanna Walter, Aaron Hill, Kathryn Ray, and Jared Smith. 2011. *Breaking the Low-Pay, No-Pay Cycle: Final Evidence from the UK Employment Retention and Advancement (ERA) Demonstration*. Research Report No 765. London, England: Department for Work and Pensions.

Hendren, Nathaniel. 2016. "The Policy Elasticity." *Tax Policy and the Economy* 30 (1): 51–89.

Hendren, Nathaniel. 2020. "Measuring Economic Efficiency Using Inverse-Optimum Weights." NBER Working Paper No. 20351. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w20351.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2019. "Unified Welfare Analysis of Government Policies." NBER WP No. 26144. https://www.nber.org/papers/w26144.

Herd, Pamela, and Donald P. Moynihan. 2018. *Administrative Burden: Policymaking by Other Means*. New York: Russell Sage Foundation.

Hernandez, Brigida, Mary J. Cometa, Jay Rosen, Jessica Velcoff, Daniel Schober, and Rene D. Luna. 2006. "Employment, Vocational Rehabilitation, and the Ticket to Work Program: Perspectives of Latinos with Disabilities." *Journal of Applied Rehabilitation Counseling* 37 (3): 13–22.

HHS/ACF/OPRE (US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation). 2020. *Portfolio of Research in Welfare and Family Self-Sufficiency*. OPRE Report 2021-13. Washington, DC: Author.

Higgins, Julian P.T., and Simon G. Thompson. 2004. "Controlling the Risk of Spurious Findings from Meta-Regression." *Statistics in Medicine* 23 (11): 1663–1682.

Hill, Fiona. 2020. "Public Service and the Federal Government." *Policy 2020 Voter Vitals.* Washington, DC: Brookings Institution.

Hirano, Kara A., Dawn Rowe, Lauren Lindstrom, and Paula Chan. 2018. "Systemic Barriers to Family Involvement in Transition Planning for Youth with Disabilities: A Qualitative Metasynthesis." *Journal of Child and Family Studies* 27 (11): 3440–3456.

Hock, Heinrich, Michael Levere, Kenneth Fortson, and David Wittenburg. 2019. *Lessons from Pilot Tests of Recruitment for the Promoting Opportunity Demonstration.* Report for Social Security Administration, Office of Research, Demonstration, and Employment Support. Washington, DC: Mathematica Policy Research.

Hock, Heinrich, Dara Lee Luca, Tim Kautz, and David Stapleton. 2017. *Improving the Outcomes of Youth with Medical Limitations through Comprehensive Training and Employment Services: Evidence from the National Job Corps Study.* Washington, DC: Mathematica Policy Research.

Hock, Heinrich, David Wittenburg, and Michael Levere. 2020. "Memorandum: Promoting Opportunity Demonstration: Recruitment and Random Assignment Report." Washington, DC: Mathematica Policy Research.

Hock, Heinrich, David Wittenburg, Michael Levere, Noelle Denny-Brown, and Heather Gordon. 2020. *Promoting Opportunity Demonstration: Recruitment and Random Assignment Report.* Washington, DC: Mathematica Policy Research.

Hoffman, Denise, Sarah Croake, David R. Mann, David Stapleton, Priyanka Anand, Chris Jones, Judy Geyer, Daniel Gubits, Stephen Bell, Andrew McGuirk, David Wittenburg, Debra Wright, Amang Sukasih, David Judkins, and Michael Sinclair. 2017. *2016 Stage 1 Interim Process, Participation, and Impact Report.* Report for the Social Security Administration (contract deliverable 24c2.1 under Contract SS00-10-60011), Office of Program Development & Research. Cambridge, MA: Abt Associates; and Washington, DC: Mathematica Policy Research.

Hoffman, Denise, Jeffrey Hemmeter, and Michelle S. Bailey. 2018. "The Relationship between Youth Services and Adult Outcomes among Former Child SSI Recipients." *Journal of Vocational Rehabilitation* 48 (2): 233–247.

Hoffmann, Holger, Dorothea Jäckel, Sybille Glauser, Kim T. Mueser, and Zeno Kupper. 2014. "Long-Term Effectiveness of Supported Employment: 5-Year Follow-Up of a Randomized Controlled Trial." *American Journal of Psychiatry* 171 (11): 1183–1190.

Holbrook, Allyson L., Timothy P. Johnson, and Maria Krysan. 2019. "Race- and Ethnicity-of-Interviewer Effects." In *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 197–224. Hoboken, NJ: John Wiley & Sons.

Hollenbeck, Kevin. 2015. *Promoting Retention or Reemployment of Workers after a Significant Injury or Illness.* Report for US Department of Labor, Office of Disability Employment Policy. Washington, DC: Mathematica Policy Research.

Hollenbeck, K. 2021. *Demonstration Evidence of Early Intervention Policies and Practices.* Kalamazoo, MI: W. E. Upjohn Institute.

Hollister, Robinson G., Peter Kemper, and Rebecca A Maynard. 1984. *The National Supported Work Demonstration.* Madison, WI: University of Wisconsin Press.

Holt, Stephen, and Katie Vinopal. 2021. "It's About Time: Examining Inequality in the Time Cost of Waiting." *SSRN*. https://doi.org/10.2139/ssrn.3857883.

Honeycutt, Todd, Kara Contreary, and Gina Livermore. 2021. *Considerations for the Papers Developed for the SSI Youth Solutions Project.* Report for the US Department of Labor, Office of Disability Employment Policy. Princeton, NJ: Mathematica. https://www.mathematica.org/publications/considerations-for-the-papers-developed-for-the-ssi-youth-solutions-project.

Honeycutt, Todd, Brittney Gionfriddo, Jacqueline Kauff, Joseph Mastrianni, Nicholas Redel, and Adele Rizzuto. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): Arkansas PROMISE Process Analysis Report.* Washington, DC: Mathematica Policy Research.

Honeycutt, Todd, Brittney Gionfriddo, and Gina Livermore. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): PROMISE Programs' Use of Effective Transition Practices in Serving Youth with Disabilities.* Washington, DC: Mathematica Policy Research.

Honeycutt, Todd, and Gina Livermore. 2018. *Promoting Readiness in Minors in Supplemental Security Income (PROMISE): The Role of PROMISE in the Landscape of Federal Programs Targeting Youth with Disabilities.* Washington, DC: Mathematica Policy Research.

Honeycutt, Todd, Eric Morris, and Thomas Fraker. 2014. *Preliminary YTD Benefit-Cost Analysis Using Administrative Data.* Princeton, NJ: Mathematica Policy Research.

Honeycutt, T., and Stapleton, D. 2013. "Striking While the Iron Is Hot: The Effect of Vocational Rehabilitation Service Wait Times on Employment Outcomes for Applicants Receiving Social Security Disability Benefits." *Journal of Vocational Rehabilitation* 39 (2): 137–152.

Honeycutt, Todd, David Wittenburg, Kelli Crane, Michael Levere, Richard Luecking, and David Stapleton. 2018. *Supplemental Security Income Youth Formative Research Project: Considerations for Identifying Promising and Testable Interventions.* Washington, DC: Mathematica Policy Research.

Honeycutt, Todd, David Wittenburg, Michael Levere, and Sarah Palmer. 2018. *Supplemental Security Income Youth Formative Research Project: Target Population Profiles.* Washington, DC: Mathematica Policy Research.

Hotz, V. Joseph, and John Karl Scholz. 2001. "Measuring Employment Income for Low-Income Populations with Administrative and Survey Data." In *Studies of Welfare Populations: Data Collection and Research Issues,* edited by M. V. Ploeg, R. A. Moffit, and C. F. Citro, 275–315. Washington, DC: The National Academies Press.

Hotz, V. J., and J. K. Scholz. 2003. "The Earned Income Tax Credit." In *Means-Tested Transfer Programs in the United States,* edited by R. Moffitt, 141–198. Chicago: University of Chicago Press.

Hoynes, H. W., and R. Moffitt. 1999. "Tax Rates and Work Incentives in the Social Security Disability Insurance Program: Current Law and Alternative Reforms." *National Tax Journal* 52 (4): 623–654.

Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2011. "Sources of Lifetime Inequality." *American Economic Review* 101 (7): 2923–2954.

Hullegie, Patrick, and Pierre Koning. 2015. "Employee Health and Employer Incentives." Discussion Paper No. 9310. Bonn, Germany: Institute for the Study of Labor.

Hussey, Michael A., and James P. Hughes. 2007. "Design and Analysis of Stepped Wedge Cluster Randomized Trials." *Contemporary Clinical Trials* 28 (2): 182–191.

IAIABC (International Association of Industrial Accident Boards and Commissions), Disability Management and Return to Work Committee. 2016. *Return to Work: A Foundational Approach to Return to Function*. Madison, WI: Author.

Ibarraran, Pablo, Laura Ripani, Bibiana Taboada, Juan Miguel Villa, and Brigida Garcia. 2014. "Life Skills, Employability, and Training for Disadvantaged Youth: Evidence from a Randomized Evaluation Design." *IZA Journal of Labor & Development* 3 (1): 1–24.

Imai, K., D. Tingley, and T. Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (1): 5-51.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001.

Imbens, Guido W., and Donald B. Rubin. 2015. *An Introduction to Causal Inference in Statistics, Biomedical and Social Sciences.* New York: Cambridge University Press.

Inanc, Hande, and David R. Mann. 2019. "Recent Changes and Reforms to the United Kingdom's Income Support Program for People with Disabilities." Center for Studying Disability Policy, Working Paper 2019-16. Washington, DC: Mathematica.

Iwanaga, Kanako, Paul Wehman, Valerie Brooke, Lauren Avellone, and Joshua Taylor. 2021. "Evaluating the Effect of Work Incentives Benefits Counseling on Employment Outcomes of Transition-Age and Young Adult Supplemental Security Income Recipients with Intellectual Disabilities: A Case Control Study." *Journal of Occupational Rehabilitation* 31: 581-591.

Johnson, George E. 1979. "The Labor Market Displacement Effect in the Analysis of the Net Impact of Manpower Training Programs." *Research in Labor Economics*, Supplement 1, 227–254.

Johnson, George E., and James D. Tomola. 1977. "The Fiscal Substitution Effect of Alternative Approaches to Public Service Employment Policy." *Journal of Human Resources* 12 (1): 3–26.

Kanter, Joel. 1989. "Clinical Case Management: Definition, Principles, Components." *Psychiatric Services* 40 (4): 361–368.

Kapteyn, Arie, and Jelmer Y. Ypma. 2007. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data." *Journal of Labor Economics* 25 (3): 513–551.

Karhan, Andrew J., and Thomas P. Golden. 2021. *Policy Considerations for Implementing Youth and Family Case Management Strategies across Systems*. Washington, DC: US Department of Labor, Office of Disability Employment Policy.

Katz, Lawrence F. 1994. "Active Labor Market Policies to Expand Employment and Opportunity." In *Reducing Unemployment: Current Issues and Policy Options*, 239–290. Jackson Hole, WY: Federal Reserve Bank of Kansas City.

Kauff, Jacqueline, Jonathan Brown, Norma Altshuler, and Noelle Denny-Brown. 2009. *Findings from a Study of the SSI/SSDI Outreach, Access, and Recovery (SOAR) Initiative.* Washington, DC: Mathematica Policy Research.

Kauff, Jacqueline F., Elizabeth Clary, Kristin Sue Lupfer, and Pamela J. Fischer. 2016. "An Evaluation of SOAR: Implementation and Outcomes of an Effort to Improve Access to SSI and SSDI." *Psychiatric Services* 67 (10): 1098–1102.

Kauff, Jacqueline, Elizabeth Clary, and Julia Lyskawa. 2014. *An Evaluation of SOAR: The Implementation and Outcomes of an Effort to Increase Access to SSI and SSDI.* Washington, DC: Mathematica Policy Research.

Kauff, Jacqueline, Todd Honeycutt, Karen Katz, Joseph Mastrianni, and Adele Rizzuto. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): Maryland PROMISE Process Analysis Report.* Washington, DC: Mathematica Policy Research.

Kennedy, Courtney, and Hannah Hartig. 2019. "Response Rates in Telephone Surveys Have Resumed Their Decline" (blog), *Pew Research Center.* February 27, 2019. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

Kennedy, Elizabeth, and Laura King. 2014. "Improving Access to Benefits for Persons with Disabilities Who Were Experiencing Homelessness: An Evaluation of the Benefits Entitlement Services Team Demonstration Project." *Social Security Bulletin* 74 (4): 45–55.

Kerachsky, Stuart, and Craig Thornton. 1987. "Findings from the STETS Transitional Employment Demonstration." *Exceptional Children* 53 (6): 515–521.

Kerachsky, Stuart, Craig Thornton, Anne Bloomenthal, Rebecca Maynard, and Susan Stephens. 1985. *Impacts of Transitional Employment on Mentally Retarded Young Adults: Results of the STETS Demonstration.* Washington, DC: Mathematica Policy Research.

Kerksick, Julie, David Riemer, and Conor Williams. 2016. "Using Transitional Jobs to Increase Employment of SSDI Applicants and Beneficiaries." In *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program*, edited by Committee for a Responsible Federal Budget, The McCrery-Pomeroy SSDI Solutions Initiative, Ch. 5. West Conshohocken, PA: Infinity Publishing.

Kimball, Miles S. 1990. "Precautionary Saving in the Small and in the Large." *Econometrica* 58 (1): 53–73.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching" *Political Analysis* 27 (4): 435–454.

Klerman, Jacob. 2020. "Findings from the (Experimental) Job Training Literature." Abt Associates. Mimeo.

Kluve, Jochen, Susana Puerto, David Robalino, Jose Maunel Romero, Friederike Rother, Jonathan Stöterau, Felix Weidenkaff, and Marc Witte. 2016. "Do Youth Employment Programs Improve Labor Market Outcomes? A Systematic Review." IZA Discussion Paper, No. 10263. Bonn, Germany: Institute for the Study of Labor. https://ftp.iza.org/dp10263.pdf.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. 2020. "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach." *Journal of Human Resources*. https://doi.org/10.3368/jhr.57.2.0718-9615R1.

Ko, Hansoo, Renata E. Howland, and Sherry A. Glied. 2020. "The Effects of Income on Children's Health: Evidence from Supplemental Security Income Eligibility under New York State Medicaid." NBER Working Paper No. 26639. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w26639.

Kogan, Deborah, Hannah Betesh, Marian Negoita, Jeffrey Salzman, Laura Paulen, Haydee Cuza, Liz Potamites, Jillian Berk, Carrie Wolfson, and Patty Cloud. 2012. *Evaluation of the Senior Community Service Employment Program (SCSEP) Process and Outcomes Study Final Report.* Report for US Department of Labor, Employment and Training Administration, Office of Policy Development and Research. Oakland, CA: Social Policy Research Associates.

Kornfeld, Robert, and Kalman Rupp. 2000. "The Net Effects of the Project NetWork Return-to-Work Case Management Experiment on Participant Earnings, Benefit Receipt, and Other Outcomes." *Social Security Bulletin* 63 (1): 12–33.

Kornfeld, Robert J., Michelle L. Wood, Larry L. Orr, and David A. Long. 1999. *Impacts of the Project NetWork Demonstration: Final Report.* Report for Social Security Administration. Bethesda, MD: Abt Associates.

Kregel, John. 2006a. *Conclusions Drawn from the State Partnership Initiative.* Richmond, VA: Virginia Commonwealth University, Rehabilitation Research and Training Center, State Partnership Systems Change Initiative Project Office. https://www.ssa.gov/disabilityresearch/documents/spiconclusions.pdf.

Kregel, John. 2006b. *Final Evaluation Report of the SSI Work Incentives Demonstration Project.* Richmond, VA: Virginia Commonwealth University, Rehabilitation Research and Training Center, State Partnership Systems Change Initiative Project Office. https://www.ssa.gov/disabilityresearch/documents/spireport.pdf.

Kregel, John, and Susan O'Mara. 2011. "Work Incentive Counseling as a Workplace Support." *Journal of Vocational Rehabilitation* 35 (2): 73–83. https://www.doi.org/10.3233/JVR-2011-0555.

Kunz, Tanja, and Marek Fuchs. 2019. "Using Experiments to Assess Interactive Feedback That Improves Response Quality in Web Surveys." In Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 247–274. Hoboken, NJ: John Wiley & Sons.

Larson, Sheryl A., and Judy Geyer. 2021. "Delaying Application of SSI's Substantial Gainful Activity Eligibility Criterion from Age 18 to 22." Washington, DC: US Department of Labor, Office of Disability Employment Policy.

Lavrakas, Paul J., Jenny Kelly, and Colleen McClain. 2019. "Investigating Interviewer Effects and Confounds in Survey-Based Experimentation." In *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 225–244. Hoboken, NJ: John Wiley & Sons.

Leiter, Valerie, Michelle L. Wood, and Stephen H. Bell. 1997. "Case Managements at Work for SSA Disability Beneficiaries: Process Results of the Project NetWork Return-to-Work Demonstration." *Social Security Bulletin* 60: 29–48.

Levere, Michael, Todd Honeycutt, Gina Livermore, Arif Mamun, and Karen Katz. 2020. *Family Service Use and Its Relationship with Youth Outcomes*. Washington, DC: Mathematica Policy Research.

Levy, Frank. 1979. "The Labor Supply of Female Household Heads, or AFDC Work Incentives Don't Work Too Well." *Journal of Human Resources* 14 (1): 76–97.

Liebman, Jeffrey B. 2015. "Understanding the Increase in Disability Insurance Benefit Receipt in the United States." *Journal of Economic Perspectives* 29 (2): 123–150.

Liebman, Jeffrey B., and Jack A. Smalligan. 2013. "Proposal 4: An Evidence-Based Path to Disability Insurance Reform." In *15 Ways to Rethink the Federal Budget*, 27–30. Washington, DC: The Hamilton Project.

Liu, Su, and David C. Stapleton. 2011. "Longitudinal Statistics on Work Activity and Use of Employment Supports for New Social Security Disability Insurance Beneficiaries." *Social Security Bulletin* 71 (3): 35–59.

Livermore, Gina. 2011. "Social Security Disability Beneficiaries with Work-Related Goals and Expectations." *Social Security Bulletin* 71 (3): 61–82.

Livermore, Gina A., and Nanette Goodman. 2009. *A Review of Recent Evaluation Efforts Associated with Programs and Policies Designed to Promote the Employment of Adults with Disabilities*. Princeton, NJ: Mathematica Policy Research.

Livermore, Gina, Todd Honeycutt, Arif Mamun, and Jacqueline Kauff. 2020. "Insights about the Transition System for SSI Youth from the National Evaluation of Promoting Readiness of Minors in SSI (PROMISE)." *Journal of Vocational Rehabilitation* 52 (1): 1–17.

Livermore, Gina, Arif Mamun, Jody Schimmel, and Sarah Prenovitz. 2013. *Executive Summary of the Seventh Ticket to Work Evaluation Report.* Washington, DC: Mathematica Policy Research.

Livermore, Gina, and Sarah Prenovitz. 2010. *Benefits Planning, Assistance, and Outreach (BPAO) Service User Characteristics and Use of Work Incentives*. Work Activity and Use of Employment Supports under the Original Ticket to Work Regulations, Final Report. No. 5ca13079097b4ae887f19a614aca2bec. Washington, DC: Mathematica Policy Research.

Livermore, Gina, David Wittenburg, and David Neumark. 2014. "Finding Alternatives to Disability Benefit Receipt." *IZA Journal of Labor Policy* 3 (14). https://doi.org/10.1186/2193-9004-3-14.

Lowenstein, Amy E., Noemi Altman, Patricia M. Chou, Kristen Faucetta, Adam Greeney, Daniel Gubits, Jorgen Harris, JoAnn Hsueh, Erika Lundquist, Charles Michalopoulos, and Vinh Q. Nguyen. 2014. *A Family-Strengthening Program for Low-Income Families: Final Impacts from the Supporting Healthy Marriage Evaluation, Technical Supplement*. OPRE Report 2014-09B. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.

Luecking, Richard G., and David C. Wittenburg. 2009. "Providing Supports to Youth with Disabilities Transitioning to Adulthood: Case Descriptions from the Youth Transition Demonstration." *Journal of Vocational Rehabilitation*, 30: 241–251.

Maestas, Nicole. 2019. "Identifying Work Capacity and Promoting Work: A Strategy for Modernizing the SSDI Program." *The ANNALS of the American Academy of Political and Social Science* 686 (1): 93–120.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review* 103 (5): 1797–1829.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. Forthcoming. "The Effect of Economic Conditions on the Disability Insurance Program: Evidence from the Great Recession." *Journal of Public Economics.*

Maestas, Nicole, Kathleen J. Mullen, and Gema Zamarro. 2010. *Research Designs for Estimating Induced Entry into the SSDI Program Resulting from a Benefit Offset.* Santa Monica, CA: The RAND Corporation.

Malani, Anup. 2006. "Identifying Placebo Effects with Data from Clinical Trials." *Journal of Political Economy* 114 (2): 236–256.

Mamun, Arif, Ankita Patnaik, Michael Levere, Gina Livermore, Todd Honeycutt, Jacqueline Kauff, Karen Katz, AnnaMaria McCutcheon, Joseph Mastrianni, and Brittney Gionfriddo. 2019. *Promoting Readiness of Minors in SSI (PROMISE) Evaluation: Interim Services and Impact Report*. Washington, DC: Mathematica Policy Research.

Mamun, Arif, David Wittenburg, Noelle Denny-Brown, Michael Levere, David R. Mann, Rebecca Coughlin, Sarah Croake, Heather Gordon, Denise Hoffman, Rachel Holzwat, Rosalind Keith, Brittany McGill, and Aleksandra Wec. 2021. *Promoting Opportunity Demonstration: Interim Evaluation Report.* Report for Social Security Administration, Office of Research, Demonstration, and Employment Support. Washington, DC: Mathematica Policy Research.

Manchester, Joyce. 2019. *Targeting Early Intervention Based on Health Care Utilization of SSDI Beneficiaries by State, with Emphasis on Mental Disorders and Substance Abuse.* Washington, DC: Committee for a Responsible Federal Budget, McCrery-Pomeroy SSDI Solutions Initiative. https://www.crfb.org/sites/default/files/Targeting_Early_Intervention_Based_On_Health_Care_Utilization.pdf.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341 (6149): 976–980.

Marrow Jocelyn, Daley Tamara, Taylor Jeffrey, Karakus Mustafa, Marshall Tina, Lewis Megan. 2020. *Supported Employment Demonstration. Interim Process Analysis Report (Deliverable 7.5a).* Rockville, MD: Westat. https://www.ssa.gov/disabilityresearch/documents/SED_Interim_Process_Analysis_Report_8-07-20.pdf.

Martin, F., and Sevak, P. 2020. "Implementation and Impacts of the Substantial Gainful Activity Project Demonstration in Kentucky." *Journal of Vocational Rehabilitation* (Preprint), 1-9.

Martin, Patricia P. 2016. "Why Researchers Now Rely on Surveys for Race Data on OASDI and SSI Programs: A Comparison of Four Major Surveys." *Research and Statistics Notes.* No. 2016-01. Social Security Administration. https://www.ssa.gov/policy/docs/rsnotes/rsn2016-01.html.

Martinez, John, Thomas Fraker, Michelle Manno, Peter Baird, Arif Mamun, Bonnie O'Day, Anu Rangarajan, David Wittenburg, and Social Security Administration. 2010. *Social Security Administration's Youth Transition Demonstration Projects: Implementation Lessons from the Original Sites*. Washington, DC: Mathematica Policy Research.

Martinson, Karin, Doug McDonald, Amy Berninger, and Kyla Wasserman. 2021. *Building Evidence-Based Strategies to Improve Employment Outcomes for Individuals with Substance Use Disorders.* OPRE Report 2020-171. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Matulewicz, Holly, Karen Katz, Todd Honeycutt, Jacqueline Kauff, Joseph Mastrianni, Adele Rizzuto, and Claire S. Wulsin. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): California PROMISE Process Analysis Report*. Washington, DC: Mathematica Policy Research.

Maximus. 2002. *Youth Continuing Disability Review Project: Annual Report October 1, 2001–September 30, 2002.* Report to the Social Security Administration, Office of Employment Support Programs.

McCann, Ted, and Nick Hart. 2019. "Disability Policy: Saving Disability Insurance with the First Reforms in a Generation." In *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*, edited by Nick Hart and Meron Yohannes, 28–39. Washington, DC: Bipartisan Policy Center.

McConnell, Sheena, and Steven Glazerman. 2001. *National Job Corps Study: The Benefits and Costs of Job Corps*. Washington, DC: Mathematica Policy Research.

McConnell, Sheena, Irma Perez-Johnson, and Jillian Berk. 2014. "Proposal 9: Providing Disadvantaged Workers with Skills to Succeed in the Labor Market." In *Policies to Address Poverty in America*, edited by Melissa S. Kearney and Benjamin H. Harris, 97–189. Washington, DC: The Brookings Institution.

McCoy, Marion L., Cynthia S. Robins, James Bethel, Carina Tornow, and William D. Frey. 2007. *Evaluation of Homeless Outreach Projects and Evaluation: Task 6: Final Evaluation Report*. Rockville, MD: Westat.

McCutcheon, AnnaMaria, Karen Katz, Rebekah Selekman, Todd Honeycutt, Jacqueline Kauff, Joseph Mastrianni, and Adele Rizzuto. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): New York State PROMISE Process Analysis Report*. Washington, DC: Mathematica Policy Research.

McHugo, G. J., R. E. Drake, R. Whitley, G. R. Bond, K. Campbell, C. A. Rapp, H. H. Goldman, W. J. Lutz, and M. T. Finnerty. 2007. "Fidelity Outcomes in the National Implementing Evidence-Based Practices Project." *Psychiatric Services* 58: 1279–1284.

McLaughlin, James R. 1994. "Estimated Increase in OASDI Benefit Payments That Would Result from Two 'Earnings Test' Type Alternatives to the Current Criteria for Cessation of Disability Benefits—Information." Memorandum, SSA Office of the Actuary.

Metcalf, C. E. 1973. "Making Inferences from Controlled Income Maintenance Experiments." *American Economic Review* 63 (3): 478–483.

Meyer, Bruce D. 1995. "Lessons from the US Unemployment Insurance Experiments." *Journal of Economic Literature* 33 (1): 91–131.

Meyers, Marcia K., Janet C. Gornick, and Laura R. Peck. 2002. "More, Less, or More of the Same? Trends in State Social Welfare Policy in the 1990s." *Publius: The Journal of Federalism* 32 (4): 91–108.

Michalopoulos, Charles, David Wittenburg, Dina A. R. Israel, Jennifer Schore, Anne Warren, Aparajita Zutshi, Stephen Freedman, and Lisa Schwartz. 2011. *The Accelerated Benefits Demonstration and Evaluation Project: Impacts on Health and Employment at Twelve Months*. New York: MDRC. http://www.ssa.gov/disabilityresearch/documents/AB%20Vol%201_508%20comply.pdf.

Miller, L., and S. O'Mara. 2003 [updated 2004]. "Social Security Disability Benefit Issues Affecting Transition Aged Youth." Briefing Paper, vol. 8. Richmond, VA: Virginia Commonwealth University, Benefits Assistance Resource Center.

Moffitt, Robert A. 1992a. "Evaluation Methods for Program Entry Effects." In *Evaluating Welfare and Training Programs*, edited by C. F. Manski and I. Garfinkel, 231–252. Cambridge, MA: Harvard University Press.

Moffitt, Robert. 1992b. "Incentive Effects of the US Welfare System: A Review." *Journal of Economic Literature* 30 (1): 1–61.

Moffitt, Robert A. 1996. "The Effect of Employment and Training Programs on Entry and Exit from the Welfare Caseload." *Journal of Policy Analysis and Management* 15 (1): 32–50.

Moffitt, Robert, ed. 2016. *Economics of Means-Tested Transfer Programs in the United States*. Chicago: University of Chicago Press.

Mojtabai, Ramin. 2011. "National Trends in Mental Health Disability, 1997–2009." *American Journal of Public Health* 101 (11): 2156–2163.

Moynihan, Donald, Pamela Herd, and Hope Harvey. 2015. "Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions." *Journal of Public Administration Research and Theory* 25 (1): 43–69.

Mullen, Kathleen J., and Stephanie L. Rennane. 2017. "The Effect of Unconditional Cash Transfers on the Return to Work of Permanently Disabled Workers." NBER Working Paper No. DRC NB17-09. Cambridge, MA: National Bureau of Economic Research: https://www.nber.org/programs-projects/projects-and-centers/retirement-and-disability-research-center/center-papers/drc-nb17-09.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2015. *Mental Disorders and Disabilities among Low-Income Children.* Washington, DC: The National Academies Press. https://doi.org/10.17226/21780.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. *Opportunities for Improving Programs and Services for Children with Disabilities.* Washington, DC: The National Academies Press.

National Association of Social Work. 2013. "NASW Standards for Social Work Case Management." https://www.socialworkers.org/LinkClick.aspx?fileticket=acrzqmEfhlo%3D&portalid=0.

National Disability Institute. 2020. *Race, Ethnicity, and Disability: The Financial Impact of Systemic Inequality and Intersectionality.* Washington, DC: Author. https://www.nationaldisabilityinstitute.org/wp-content/uploads/2020/08/race-ethnicity-and-disability-financial-impact.pdf.

National Safety Council. 2020. "NSC Injury Facts." https://injuryfacts.nsc.org/.

Nazarov, Zafar. 2013. "Can Benefits and Work Incentives Counseling Be a Path to Future Economic Self-Sufficiency for SSI/SSDI Beneficiaries?" Working Paper No. 2013-17. Chestnut Hill, MA: Center for Retirement Research at Boston College.

NCWD/Y (National Collaborative on Workforce and Disability for Youth). 2005. *Guideposts for Success.* Washington, DC: Institute on Education Leadership, 2005.

NCWD/Y (National Collaborative on Workforce and Disability for Youth). 2009. *Guideposts for Success,* 2nd ed. Washington, DC: Institute on Educational Leadership.

NCWD/Y (National Collaborative on Workforce and Disability for Youth). 2019. *Guideposts for Success 2.0: A Framework for Successful Youth Transition to Adulthood.* Washington, DC: Author. http://www.ncwd-youth.info/wp-content/uploads/2019/07/Guideposts-for-Success-2.0.pdf.

Neuhauser, Frank. 2016, April. "The Myth of Workplace Injuries: Or Why We Should Eliminate Workers' Compensation for 90% of Workers and Employers." *IAIABC Perspectives*. https://resources.iaiabc.org/1a4arng/.

Nichols, Austin, Emily Dastrup, Zachary Epstein, and Michelle Wood. 2020. *Data Analysis for Stay-at-Work/Return-to-Work (SAW/RTW) Models and Strategies Project. Early Intervention Pathway Map and Population Profiles.* Report for US Department of Labor. Cambridge, MA: Abt Associates.

Nichols, A., J. Geyer, M. Grosz, Z. Epstein, and M. Wood. 2020. *Synthesis of Evidence about Stay-at-Work/ Return-to-Work (SAW/RTW) and Related Programs.* Report for the U.S. Department of Labor. Rockville, MD: Abt Associates.

Nichols, Austin, and Jesse Rothstein. 2016. "The Earned Income Tax Credit." In *Economics of Means-Tested Transfer Programs in the United States*, Vol. 1, edited by Robert A. Moffitt, 137–218. Chicago: University of Chicago Press.

Nichols, Austin, Lucie Schmidt, and Purvi Sevak. 2017. "Economic Conditions and Supplemental Security Income Applications." *Social Security Bulletin* 77 (4): 27–44.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. "The Impressive Effects of Tutoring on Prek–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence." NBER Working Paper No. 27476. Cambridge, MA: National Bureau of Economic Research.

Noel, Valerie A., Eugene Oulvey, Robert E. Drake, Gary R. Bond, Elizabeth A. Carpenter-Song, and Brian DeAtley. 2018. "A Preliminary Evaluation of Individual Placement and Support for Youth with Developmental and Psychiatric Disabilities." *Journal of Vocational Rehabilitation* 48 (2): 249–255.

NTACT (National Technical Assistance Center on Transition). 2016. *Evidence-Based Practices and Predictors in Secondary Transition: What We Know and What We Still Need to Know*. Charlotte, NC: Author. https://transitionta.org/wp-content/uploads/docs/EBPP_Exec_Summary_2016_12-13.pdf.

Nunn, Ryan, Jana Parsons, and Jay Shambaugh. 2019. *Labor Force Nonparticipation: Trends, Causes, and Policy Solutions*. The Hamilton Project. Washington, DC: Brookings. https://www.hamiltonproject.org/assets/files/PP_LFPR_final.pdf.

Nye-Lengerman, Kelly, Amy Gunty, David Johnson, and Maureen Hawes. 2019. "What Matters: Lessons Learned from the Implementation of PROMISE Model Demonstration Projects." *Journal of Vocational Rehabilitation* 51 (2): 275–284.

O'Day, Bonnie, Hannah Burak, Kathleen Feeney, Elizabeth Kelley, Frank Martin, Gina Freeman, Grace Lim, and Katie Morrison. 2016. *Employment Experiences of Young Adults and High Earners Who Receive Social Security Disability Benefits: Findings from Semistructured Interviews*. Washington, DC: Mathematica Policy Research.

O'Day, Bonnie, Allison Roche, Norma Altshuler, Liz Clary, and Krista Harrison. 2009. *Process Evaluation of the Work Incentives Planning and Assistance Program.* Work Activity and Use of Employment Supports under the Original Ticket to Work Regulations, Report 1. Washington, DC: Mathematica Policy Research.

O'Leary, Paul, Leslie I. Boden, Seth A. Seabury, Al Ozonoff, and Ethan Scherer. 2012. "Workplace Injuries and the Take-Up of Social Security Disability Benefits." *Social Security Bulletin* 72 (3): 1–17.

Olney, Marjorie F., and Cindy Lyle. 2011. "The Benefits Trap: Barriers to Employment Experienced by SSA Beneficiaries." *Rehabilitation Counseling Bulletin* 54 (4): 197–209.

Olsen, Anya, and Russell Hudson. 2009. "Social Security Administration's Master Earnings File: Background Information," *Social Security Bulletin* 69 (3): 29–46.

Olsen, Robert B., Larry L. Orr, Stephen H. Bell, and Elizabeth A. Stuart. 2013. "External Validity in Policy Evaluations That Choose Sites Purposively." *Journal of Policy Analysis and Management* 32 (1): 107–121. https://doi.org/10.1002/pam.21660.

Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage.

Page, Lindsay C., Avi Feller, Todd Grindal, Luke Miratrix, and Marie-Andree Somers. 2015. "Principal Stratification: A Tool for Understanding Variation in Program Effects across Endogenous Subgroups." *American Journal of Evaluation* 36 (4): 514–531.

Parsons, Donald O. 1980. "The Decline in Male Labor Force Participation." *Journal of Political Economy* 88 (1): 117–134.

Peck, Laura R. 2003. "Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post Treatment Choice." *American Journal of Evaluation* 24 (2): 157–187.

Peck, Laura R. 2005. "Using Cluster Analysis in Program Evaluation." *Evaluation Review* 29: (25): 178–196.

Peck, Laura R. 2013. "On Analysis of Symmetrically Predicted Endogenous Subgroups: Part One of a Method Note in Three Parts." *American Journal of Evaluation* 34 (2): 225–236.

Peck, Laura R. 2020. *Experimental Evaluation Design for Program Improvement*. Thousand Oaks, CA: Sage.

Peck, Laura R., Daniel Litwok, Douglas Walton, Eleanor Harvill, and Alan Werner. 2019. *Health Profession Opportunity Grants (HPOG 1.0) Impact Study: Three-Year Impacts Report.* OPRE Report 2019-114. Report for US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. Rockville, MD: Abt Associates.

Peck, Laura R., and Ronald J. Scott, Jr. 2005. "Can Welfare Case Management Increase Employment? Evidence from a Pilot Program Evaluation." *Policy Studies Journal* 33 (4): 509–533.

Peikes, Deborah N., Lorenzo Moreno, and Sean Michael Orzol. 2008. "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs." *The American Statistician* 62 (3): 222–231.

Peikes, Deborah, Sean Orzol, Lorenzo Moreno, and Nora Paxton. 2005. *State Partnership Initiative: Selection of Comparison Groups for the Evaluation and Selected Impact Estimates: Final Report*. Princeton, NJ: Mathematica Policy Research.

The Policy Surveillance Program. n.d. "State Supplemental Payments for Children with Disabilities." Accessed September 20, 2021. http://www.lawatlas.org/datasets/supplemental-security-income-for-children-with-disabilities.

Porter, Alice, James Smith, Alydia Payette, Tim Tremblay, and Peter Burt. 2009. *SSDI $1 for $1 Benefit Offset Pilot Demonstration Vermont Pilot Final Report.* Burlington, VT: Vermont Division of Vocational Rehabilitation. https://www.ssa.gov/disabilityresearch/documents/Vt1for2FinalReport091223.pdf.

Prero, Aaron J., and Craig Thornton. 1991. "Transitional Employment Training for SSI Recipients with Mental Retardation." *Social Security Bulletin* 54 (11): 2–25.

Proudlock, S., and N. Wellman. 2011. "Solution Focused Groups: The Results Look Promising." *Counselling Psychology Review* 26 (3): 45–54.

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price. 2009. "What to Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: US Department of Education.

Rangarajan, Anu, Thomas Fraker, Todd Honeycutt, Arif Mamun, John Martinez, Bonnie O'Day, and David Wittenburg. 2009. *The Social Security Administration's Youth Transition Demonstration Projects: Evaluation Design Report*. No. dc181046c9a041e6b63bb1b5743e1935. Princeton, NJ: Mathematica Policy Research.

Rothstein, Jesse, and Till von Wachter. 2017. "Social Experiments in the Labor Market." In *Handbook of Economic Field Experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 555–637. Amsterdam, The Netherlands: North-Holland/Elsevier.

Ruiz-Quintanilla, S. Antonio, Robert R. Weathers II, Valerie Melburg, Kimberly Campbell, and Nawaf Madi. 2006. "Participation in Programs Designed to Improve Employment Outcomes for Persons with Psychiatric Disabilities: Evidence from the New York WORKS Demonstration Project." *Social Security Bulletin* 66 (2): 49–79.

Rupp, Kalman, Stephen H. Bell, and Leo A. McManus. 1994. "Design of the Project NetWork Return-to-Work Experiment for Persons with Disabilities." *Social Security Bulletin* 57: 3. (2): 3–20. https://pubmed.ncbi.nlm.nih.gov/7974091/.

Rupp, Kalman, Michelle Wood, and Stephen H. Bell. 1996. "Targeting People with Severe Disabilities for Return-to-Work: The Project NetWork Demonstration Experience." *Journal of Vocational Rehabilitation* 7 (1–2): 63–91.

SAMHSA (Substance Abuse and Mental Health Services Administration). n.d. "SSI/SSDI Outreach, Access and Recovery: An Overview." Rockville, MD: Author.
https://soarworks.samhsa.gov/sites/soarworks.prainc.com/files/SOAROverview-2020-508_0.pdf.

Sampson, James P., Robert C. Reardon, Gary W. Peterson, and Janet G. Lenz. 2004. *Career Counseling and Services: A Cognitive Information Processing Approach*. Belmont, CA: Thomson/Brooks/Cole.

Schiller, Bradley R. 1973. "Empirical Studies of Welfare Dependency: A Survey." *Journal of Human Resources* 8: 19–32.

Schimmel, Jody, David Stapleton, David Mann, and Dawn Phelps. 2013. *Participant and Provider Outcomes since the Inception of Ticket to Work and the Effects of the 2008 Regulatory Changes*. Report for Social Security Administration, Office of Research, Demonstration, and Employment Support. Washington, DC: Mathematica Policy Research.

Schimmel, Jody, David C. Stapleton, and Jae G. Song. 2011. "How Common Is Parking among Social Security Disability Insurance Beneficiaries. Evidence from the 1999 Change in the Earnings Level of Substantial Gainful Activity." *Social Security Bulletin* 71 (4): 77–92.

Schlegelmilch, Amanda, Matthew Roskowski, Cayte Anderson, Ellie Hartman, and Heidi Decker-Maurer. 2019. "The Impact of Work Incentives Benefits Counseling on Employment Outcomes of Transition-Age Youth Receiving Supplemental Security Income (SSI) Benefits." *Journal of Vocational Rehabilitation* 51 (2): 127–136.

Schmidt, Lucie, and Purvi Sevak. 2004. "AFDC, SSI, and Welfare Reform Aggressiveness." *Journal of Human Resources* 39 (3): 792–812.

Schmidt, Lucie, and Purvi Sevak. 2017. "Child Participation in Supplemental Security Income: Cross- and within-State Determinants of Caseload Growth." *Journal of Disability Policy Studies* 28 (3): 131–140.

Schmidt, Lucie, Lara D. Shore-Sheppard, and Tara Watson. 2020. "The Impact of the ACA Medicaid Expansion on Disability Program Applications." *American Journal of Health Economics* 6 (4): 444–476.

Schochet, Peter Z. 2009. "An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations." *Evaluation Review* 33 (6): 539–567.

Schochet, Peter Z., John Burghardt, and Sheena McConnell. 2006. *National Job Corps Study and Longer-Term Follow-Up Study: Impact and Benefit-Cost Findings Using Survey and Summary Earnings Records Data. Final Report*. Princeton, NJ: Mathematica Policy Research.

Schochet, Peter Z., Sheena M. McConnell, and John A. Burghardt. 2003. *National Job Corps Study: Findings Using Administrative Earnings Records Data*. Princeton, NJ: Mathematica Policy Research, Inc.

Selekman, Rebekah, Mary A. Anderson, Todd Honeycutt, Karen Katz, Jacqueline Kauff, Joseph Mastrianni, and Adele Rizzuto. 2018. *Promoting Readiness of Minors in Supplemental Security Income (PROMISE): Wisconsin PROMISE Process Analysis Report*. Washington, DC: Mathematica Policy Research.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth/Cengage Learning.

Skidmore, Sara, Debra Wright, Kirsten Barrett, and Eric Grau. 2017. *National Beneficiary Survey–General Waves Round 5.* Vol. 2: *Data Cleaning and Identification of Data Problems.* Washington, DC: Mathematica.

Smalligan, Jack, and Chantel Boyens. 2019. "Improving the Social Security Disability Determination Process." Washington, DC: Urban Institute.

Smalligan, Jack, and Chantel Boyens. 2020. "Two Proposals to Strengthen Paid-Leave Programs." Washington, DC: Urban Institute.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Social Security Advisory Board. 2016. "Representative Payees: A Call to Action." *Issue Brief.* https://www.ssab.gov/research/representative-payees-a-call-to-action/.

Solomon, Phyllis. 1992. "The Efficacy of Case Management Services for Severely Mentally Disabled Clients." *Community Mental Health Journal* 28 (3): 163–180.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–316.

SRI International. 1983. *Final Report of the Seattle-Denver Income Maintenance Experiment.* Vol. 1, *Design and Results.* Washington, DC: Government Printing Office.

SSA (Social Security Administration). 2001. "Childhood Disability: Supplemental Security Income Program. A Guide for Physicians and Other Health Care Professionals." Social Security Administration. https://www.ssa.gov/disability/professionals/childhoodssi-pub048.htm.

SSA (Social Security Administration). 2006. "Cooperative Agreements for Work Incentives Planning and Assistance Projects; Program Announcement No. SSA-OESP-06-1." *Federal Register*. https://www.federalregister.gov/documents/2006/05/16/06-4507/program-cooperative-agreements-for-work-incentives-planning-and-assistance-projects-program.

SSA (Social Security Administration). 2016. *The Social Security Administration's Plan to Achieve Self-Support Program.* Audit Report A-08-16-50030. Office of the Inspector General. https://oig-files.ssa.gov/audits/full/A-08-16-50030.pdf.

SSA (Social Security Administration). 2018a. *National Beneficiary Survey: Disability Statistics, 2015*. Baltimore, MD: Author.

SSA (Social Security Administration). 2018b. *Social Security Programs throughout the World: Europe, 2018.* SSA Publication No. 13-11801. Washington, DC: Social Security Administration, Office of Research, Evaluation, and Statistics, Office of Retirement and Disability Policy.

SSA (Social Security Administration). 2019a. *Annual Report on Medical Continuing Reviews: Fiscal Year 2015*. Baltimore, MD: Author. https://www.ssa.gov/legislation/FY%202015%20CDR%20Report.pdf.

SSA (Social Security Administration). 2019b. *Annual Report on Section 234 Demonstration Projects.* Washington, DC: Author. https://www.ssa.gov/disabilityresearch/documents/Section%20234%20Report%20-%202019.pdf.

SSA (Social Security Administration). 2019c. *Annual Statistical Report on the Social Security Disability Insurance Program, 2018.* Washington, DC: Author. https://www.ssa.gov/policy/docs/statcomps/di_asr/2018/di_asr18.pdf.

SSA (Social Security Administration). 2019d. "Supplemental Security Income, Table 7.B1." Annual Statistical Supplement. http://www.ssa.gov/policy/docs/statcomps/supplement/2019/7b.html#table7.b1.

SSA (Social Security Administration). 2020a. *Annual Report on Section 234 Demonstration Projects*. Baltimore, MD: Author. https://www.ssa.gov/legislation/Demo%20Project%20Report%20Released%20-%20Section%20234%20Report%202020.pdf.

SSA (Social Security Administration). 2020b. *Annual Statistical Report on the Social Security Disability Insurance Program, 2019.* https://www.ssa.gov/policy/docs/statcomps/di_asr/2019/di_asr19.pdf.

SSA (Social Security Administration). 2020c. *Annual Statistical Supplement to the Social Security Bulletin*. Baltimore, MD: Author.

SSA (Social Security Administration). 2020d. *DI & SSI Program Participants: Characteristics & Employment, 2015*. Washington, DC: Author. https://www.ssa.gov/policy/docs/chartbooks/di-ssi-employment/2015/dsppce-2015.pdf.

SSA (Social Security Administration). 2020e. *Red Book. A Summary Guide to Employment Supports for People with Disabilities under the Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) Programs.* https://www.ssa.gov/redbook/.

SSA (Social Security Administration). 2020f, September. *Social Security Administration Evaluation Policy.* Washington, DC: Author. https://www.ssa.gov/data/data_governance_board/Evidence%20Act%20Evaluation%20Policy%20-%20September%202020.pdf.

SSA (Social Security Administration). 2020g. *SSA Budget Information.* https://www.ssa.gov/budget/FY21Files/2021BO.pdf.

SSA (Social Security Administration). 2020h. *SSI Annual Statistical Report, 2019*. Washington, DC: Author. https://www.ssa.gov/policy/docs/statcomps/ssi_asr/2019/ssi_asr19.pdf.

SSA (Social Security Administration). 2020i. *What You Need to Know about Your Supplemental Security Income (SSI) When You Turn 18*. Report No. 2020. Baltimore, MD: Author. www.socialsecurity.gov/pubs/EN-05-11005.pdf.

SSA (Social Security Administration). 2021. "SSI Monthly Statistics, 2020." Research, Statistics & Policy Analysis. https://www.ssa.gov/policy/docs/statcomps/ssi_monthly/2020/index.html.

SSA (Social Security Administration). n.d. "Requesting an Electronic Data Exchange with SSA." Accessed March 26, 2021. https://www.ssa.gov/dataexchange/request_dx.html.

SSA (Social Security Administration). n.d. "State Vocational Rehabilitation Agency Reimbursements." VR Reimbursement Claims Processing website. https://www.ssa.gov/work/claimsprocessing.html (accessed May 7, 2021).

SSA (Social Security Administration). n.d. "Ticket Tracker, August 2020." Accessed March 4, 2021. https://www.ssa.gov/work/tickettracker.html.

SSA/ORDP/ORDES (Social Security Administration; Office of Retirement and Disability Policy; Office of Research, Demonstration, and Employment Support). 2020. *Overview and Documentation of the Social Security Administration's Disability Analysis File (DAF) Public Use File for 2019.* Washington, DC: Mathematica. Retrieved from https://www.ssa.gov/disabilityresearch/daf_puf.html#documentation.

Stapleton, David C., Stephen H. Bell, Denise Hoffman, and Michelle Wood. 2020. "Comparison of Population-Representative and Volunteer Experiments: Lessons from the Social Security Administration's Benefit Offset National Demonstration (BOND)." *American Journal of Evaluation* 41 (4): 547–563.

Stapleton, David, Stephen Bell, David Wittenburg, Brian Sokol, and Debi McInnis. 2010. *BOND Implementation and Evaluation: BOND Final Design Report*. Report for Social Security Administration. Washington, DC: Abt Associates.

Stapleton, David, Yonatan Ben-Shalom, and David Mann. 2016. "The Employment/Eligibility System: A New Gateway for Employment Supports and Social Security Disability Benefits." In *SSDI Solutions: Ideas to Strengthen the Social Security Disability Insurance Program*, edited by Committee for a Responsible Federal Budget, The McCrery-Pomeroy SSDI Solutions Initiative, Ch. 3. Offprint. https://www.crfb.org/sites/default/files/ stapletonbenshalommann.pdf.

Stapleton, David, Yonatan Ben-Shalom, and David R. Mann. 2019. *Development of an Employment/Eligibility Services (EES) System.* Report for University of New Hampshire. Washington, DC: Mathematica Policy Research.

Stapleton, David, Robert Burns, Benjamin Doornink, Mary Harris, Robert Anfield, Winthrop Cashdollar, Brian Gifford, and Kevin Ufier. 2015. *Targeting Early Intervention to Workers Who Need Help to Stay in the Labor Force.* Report for US Department of Labor, Office of Disability Employment Policy. Washington, DC: Mathematica Policy Research.

Stapleton, David, Arif Mamun, and Jeremy Page. 2014. "Initial Impacts of the Ticket to Work Program: Estimates Based on Exogenous Variation in Ticket Mail Months." *IZA Journal of Labor Policy* 3 (1): 1–24.

State of Connecticut. 2009. *Benefit Offset Pilot Demonstration: Connecticut Final Report*. Report for Social Security Administration. https://www.ssa.gov/disabilityresearch/documents/Conn-FINAL%20BOP%20REPORT%2012%207%2009.doc.

Stepner, Michael. 2019. "The Long-Term Externalities of Short-Term Disability Insurance." Unpublished working paper. https://files.michaelstepner.com/short_term_di_externalities.pdf.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2): 369–386.

Taylor, Jeffrey, David Salkever, William Frey, Jarnee Riley, and Jocelyn Marrow. 2020. *Supported Employment Demonstration Final Enrollment Analysis Report (Deliverable 7.4b)*. Report for Social Security Administration. Rockville, MD: Westat.

Test, David W., Valerie L. Mazzotti, April L. Mustian, Catherine H. Fowler, Larry Kortering, and Paula Kohler. 2009. "Evidence-Based Secondary Transition Predictors for Improving Postschool Outcomes for Students with Disabilities." *Career Development for Exceptional Individuals* 32 (3): 160–181.

Thornton, Craig, and Paul Decker. 1989. *The Transitional Employment Training Demonstration: Analysis of Program Impacts*. Princeton, NJ: Mathematica Policy Research.

Thornton, Craig, Shari Miller Dunstan, and Jennifer Schore. 1988. *The Transitional Employment and Training Demonstration: Analysis of Program Operations*. Princeton, NJ: Mathematica Policy Research.

Thornton, Craig, Gina Livermore, Thomas Fraker, David Stapleton, Bonnie O'Day, David Wittenburg, Robert Weathers II, et al. 2007. *Evaluation of the Ticket to Work Program: Assessment of Post-Rollout Implementation and Early Impacts*, Vol. 1. Washington, DC: Mathematica Policy Research.

Tipton, Elizabeth. 2013. "Improving Generalizations from Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts" *Journal of Educational and Behavioral Statistics* 38 (3): 239–266.

Tipton, Elizabeth. 2014. "How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations." *Journal of Educational and Behavioral Statistics* 39 (6): 478–501.

Tipton, Elizabeth, and Laura R. Peck. 2017. "A Design-Based Approach to Improve External Validity in Welfare Policy Evaluations." *Evaluation Review* 41 (4): 326–356.

Tipton, Elizabeth, David S. Yeager, Ronaldo Iachan, and Barbara Schneider. 2019. "Designing Probability Samples to Study Treatment Effect Heterogeneity." In *Experimental Methods in Survey Research: Techniques That Combine Random Sampling with Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 435–456. Hoboken, NJ: John Wiley & Sons.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96 (5): 1384–1417.

Tremblay, Tim, James Smith, Alice Porter, and Robert Weathers. 2011. "Effects on Beneficiary Employment and Earnings of a Graduated $1-for-$2 Benefit Offset for Social Security Disability Insurance (SSDI)." *Journal of Rehabilitation* 77 (2): 19.

Tremblay, T., J. Smith, H. Xie, and R. Drake. 2004. "The Impact of Specialized Benefits Counseling Services on Social Security Administration Disability Beneficiaries in Vermont." *Journal of Rehabilitation* 70 (2): 5-11.

Tremblay, Timothy, James Smith, Haiyi Xie, and Robert E. Drake. 2006. "Effect of Benefits Counseling Services on Employment Outcomes for People with Psychiatric Disabilities." *Psychiatric Services* 57 (6): 816–821.

Trepper, Terry S., Yvonne Dolan, Eric E. McCollum, and Thorana Nelson. 2006. "Steve De Shazer and the Future of Solution-Focused Therapy." *Journal of Marital and Family Therapy* 32 (2): 133–139.

Treskon, Louisa. 2016. "What Works for Disconnected Young People: A Scan of the Evidence." MDRC Working Paper. New York: MDRC.

Tuma, Nancy B. 2001. "Approaches to Evaluating Induced Entry into a New SSDI Program with a $1 Reduction in Benefits for Each $2 in Earnings." Working draft prepared for the Social Security Administration. https://www.ssa.gov/disabilityresearch/documents/ind_entry_110501.pdf.

Vachon, Mallory. 2014. "The Impact of Local Labor Market Conditions and the Federal Disability Insurance Program: New Evidence from the Bakken Oil Boom." Paper presented at the 2014 Conference of the National Tax Association, Santa Fe, NM, November 2014. https://www.ntanet.org/wp-content/uploads/proceedings/2014/052-vachon-impact-local-market-conditions-federal.pdf.

Van Noorden, Richard, Brendan Maher, and Regina Nuzzo. 2014. "The Top 100 Papers." *Nature* 514 (7524): 550–553.

VanderWeele, Tyler J. 2011. "Principal Stratification—Uses and Limitations." *International Journal of Biostatistics* 7 (1): 1–14.

Vogl, Susanne, Jennifer A. Parsons, Linda K. Owens, and Paul J. Lavrakas. 2019. "Experiments on the Effects of Advance Letters in Surveys." In *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, edited by Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West, 89–110. Hoboken, NJ: John Wiley & Sons.

von Wachter, Till, Jae Song, and Joyce Manchester. 2011. "Trends in Employment and Earnings of Allowed and Rejected Applicants to the Social Security Disability Insurance Program." *American Economic Review* 101 (7): 3308–3329.

Vought, Russell T. 2020. *Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices*. Memo M-20-12. Washington, DC: Office of Management and Budget, Executive Office of the President.

Weathers II, R. R., and J. Hemmeter. 2011. "The Impact of Changing Financial Work Incentives on the Earnings of Social Security Disability Insurance (SSDI) Beneficiaries." *Journal of Policy Analysis and Management* 30 (4): 708–728.

Weathers II, Robert R., Chris Silanskis, Michelle Stegman, John Jones, and Susan Kalasunas. 2010. "Expanding Access to Health Care for Social Security Disability Insurance Beneficiaries: Early Findings from the Accelerated Benefits Demonstration." *Social Security Bulletin* 70 (4): 25–47. https://www.ssa.gov/policy/docs/ssb/v70n4/v70n4p25.html.

Weathers II, Robert R., and Michelle Stegman. 2012. "The Effect of Expanding Access to Health Insurance on the Health and Mortality of Social Security Disability Insurance Beneficiaries." *Journal of Health Economics* 31 (6): 863–875.

Weathers II, Robert R., and Michelle Stegman Bailey. 2014. "The Impact of Rehabilitation and Counseling Services on the Labor Market Activity of Social Security Disability Insurance (SSDI) Beneficiaries." *Journal of Policy Analysis and Management* 33 (3): 623–648.

Wehman, Paul H., Carol M. Schall, Jennifer McDonough, John Kregel, Valerie Brooke, Alissa Molinelli, Whitney Ham, Carolyn W. Graham, J. E. Riehle, and Holly T. Collins. 2014. "Competitive Employment for Youth with Autism Spectrum Disorders: Early Results from a Randomized Clinical Trial." *Journal of Autism and Developmental Disorders* 44 (3): 487–500.

Wehmeyer, Michael L. 1995. *The Arc's Self-Determination Scale: Procedural Guidelines*. Washington, DC: US Department of Education, Office of Special Education and Rehabilitative Services, Division of Innovation and Development.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley & Sons.

Whalen, Denise, Gilbert Gimm, Henry Ireys, Boyd Gilman, and Sarah Croake. 2012. *Demonstration to Maintain Independence and Employment (DMIE)*. Report for Centers for Medicare & Medicaid Services. Washington, DC: Mathematica Policy Research.

Wilde, Elizabeth Ty, and Robinson Hollister. 2007. "How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management* 26 (3): 455–477.

Wilhelm, Sarah, and Sara McCormick. 2013. "The Impact of a Written Benefits Analysis by Utah Benefit Counseling/WIPA Program on Vocational Rehabilitation Outcomes." *Journal of Vocational Rehabilitation* 39 (3): 219–228.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39: 453–469.

Wiseman, Michael. 2016. *Rethinking the Promoting Opportunity Demonstration Project.* Washington, DC: Social Security Advisory Board.

Wittenburg, David. 2011. *Testimony for Hearing on Supplemental Security Income Benefits for Children. Subcommittee on Human Resources, Committee on Ways and Means, US House of Representatives*. Washington, DC: Mathematica Policy Research.

Wittenburg, David, Kenneth Fortson, David Stapleton, Noelle Denny-Brown, Rosalind Keith, David R. Mann, Heinrich Hock, and Heather Gordon. 2018. *Promoting Opportunity Demonstration: Design Report*. Washington, DC: Mathematica Policy Research.

Wittenburg, David, Thomas Fraker, David Stapleton, Craig Thornton, Jesse Gregory, and Arif Mamun. 2007. "Initial Impacts of the Ticket to Work Program on Social Security Disability Beneficiary Service Enrollment, Earnings, and Benefits." *Journal of Vocational Rehabilitation* 27 (2): 129–140.

Wittenburg, David, and Gina Livermore. 2020. *Youth Transition*. Washington, DC: Mathematica Policy Research.

Wittenburg, David, David R. Mann, and Allison Thompkins. 2013. "The Disability System and Programs to Promote Employment for People with Disabilities." *IZA Journal of Labor Policy* 2 (4): 1–25.

Wittenburg, David, David Stapleton, Michelle Derr, Denise W. Hoffman, and David R. Mann. 2012. *BOND Stage 1 Early Assessment Report.* Report for Social Security Administration, Office of Research, Demonstration, and Employment Support. Cambridge, MA: Abt Associates.

Wittenburg, David, John Tambornino, Elizabeth Brown, Gretchen Rowe, Mason DeCamillis, and Gilbert Crouse. 2015. *The Child SSI Program and the Changing Safety Net*. Washington, DC: US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Office of Human Services Policy.

Wixon, Bernard, and Alexander Strand. 2013. "Identifying SSA's Sequential Disability Determination Steps Using Administrative Data." *Research and Statistics Notes.* No. 2013-01. Social Security Administration. https://www.ssa.gov/policy/docs/rsnotes/rsn2013-01.html.

youth.gov. n.d. "Job Corps, Program Activities/Goals." Accessed March 24, 2021. https://youth.gov/content/job-corps.

Zhang, C. Yiwei, Jeffrey Hemmeter, Judd B. Kessler, Robert D. Metcalfe, and Robert Weathers. 2020. "Nudging Timely Wage Reporting: Field Experimental Evidence from the United States Social Supplementary Income Program." NBER Working Paper No. 2785. Cambridge, MA: National Bureau of Economic Research.

Ziguras, Stephen J., and Geoffrey W. Stuart. 2000. "A Meta-Analysis of the Effectiveness of Mental Health Case Management over 20 Years." *Psychiatric Services* 51 (11): 1410–1421.

Contributors

Leslynn R. Angel *(Comment on Chapter 8),* President and CEO, Michigan United Cerebral Palsy—Ms. Angel has worked successfully at assisting individuals with significant disabilities find employment, utilizing customized employment. Her experience includes training Vocational Rehabilitation counselors and others and introducing choice-based philosophies and technology into the Vocational Rehabilitation system.

Burt S. Barnow *(Chapter 2),* Amsterdam Professor of Public Service and Economics, The George Washington University—Dr. Barnow teaches a doctoral program seminar on public finance and human capital. His research focuses on evaluations of workforce programs and other social programs.

Elizabeth H. Curda *(Comment on Chapter 3),* Director, Education, Workforce, and Income Security Team, US Government Accountability Office (GAO)—She oversees a portfolio of audits of federal disability programs at the Department of Veterans Affairs, the Social Security Administration, and the Railroad Retirement Board, among other agencies. Her portfolio also addresses the role federal programs play in providing equal opportunity for individuals with disabilities in all areas of public life.

Manasi Deshpande *(Comment on Chapter 6),* Assistant Professor of Economics, The University of Chicago; and Faculty Research Fellow, National Bureau of Economic Research (NBER)—Dr. Deshpande's research includes empirical public finance and labor economics, with a focus on the effects of social insurance and public assistance programs and their interaction with labor markets.

Jonah B. Gelbach *(Comment on Chapter 3),* Professor of Law, University of California, Berkeley—Dr. Gelbach's interests include civil procedure, evidence, statutory interpretation, law and economics, event study methodology, securities litigation, the economics of crime, applied statistical methodology, evaluation of public assistance programs, and general applied microeconomics.

Debra Goetz Engler *(Editor, Chapter 9),* Supervisory Social Insurance Specialist, Office of Research, Demonstration, and Employment Support, Social Security Administration (SSA)—Ms. Goetz Engler helps design, implement, and oversee research and demonstration projects related to disability and return to work.

Howard H. Goldman *(Comment on Chapter 7),* Professor of Psychiatry, University of Maryland School of Medicine—For the past decade, Dr. Goldman has chaired the National Academies of Sciences, Engineering, and Medicine, Standing Committee of Medical and Vocational Experts Assisting the Social Security Administration (SSA) on Disability Issues.

David H. Greenberg *(Chapter 2),* Professor Emeritus of Economics, University of Maryland, Baltimore County—Dr. Greenberg is a labor economist and cost-benefit analyst. Much of his research focuses on the evaluation of government programs.

Jesse Gregory *(Chapter 4),* Associate Professor, Department of Economics, University of Wisconsin-Madison—Dr. Gregory conducts research in labor, public, and urban economics. His research often focuses on questions related to the design of policies whose rules may affect individuals' location choices, including disaster relief policy, low-income housing policies, and place-based hiring subsidies.

Nick Hart *(Comment on Chapter 7),* President, The Data Foundation—Dr. Hart leads a Washington, DC-based non-profit think tank that works to improve government and society by encouraging the use of data to improve public policymaking. Dr. Hart has worked on a wide range of issues including Social Security, disability, anti-poverty, environmental, energy, economic development, and criminal justice policies.

Jeffrey Hemmeter (*Editor, Chapter 1),* Acting Deputy Associate Commissioner, Office of Research, Demonstration, and Employment Support, Social Security Administration (SSA)—Dr. Hemmeter helps design, conduct, and oversee research, evaluation, and policies related to disability and return to work. His research focuses on transition-age youth and SSI and he has worked on numerous SSA demonstrations and studies.

Kevin Hollenbeck *(Chapter 5),* Independent Economic Consultant, W.E. Upjohn Institute for Employment Research—Dr. Hollenbeck specializes in evaluations of and estimation of return on investment to education, disability, and workforce development policies.

Hilary Hoynes *(Comment on Chapter 4),* Professor of Public Policy and Economics, Goldman School of Public Policy, University of California, Berkeley—In addition to her appointment at GSPP, Dr. Hoynes holds the Haas Distinguished Chair in Economic Disparities and co-directs the Berkeley Opportunity Lab. Her research focuses on poverty, inequality, food and nutrition programs, and the impacts of government tax and transfer programs on low-income families.

Calvin Johnson *(Comment on Chapter 9),* Deputy Assistant Secretary for Research, Evaluation, and Monitoring, Office of Policy Development and Research, US Department of Housing and Urban Development (HUD)—Dr. Johnson oversees a broad evaluation portfolio to include projects exploring the impact of housing on the non-housing outcomes of economic self-sufficiency, health and wellness, and educational achievement.

John Kregel *(Comment on Chapter 8),* Professor of Special Education and Disability Policy, Virginia Commonwealth University—Dr. Kregel currently serves as the research director at the VCU Rehabilitation Research and Training Center and is co-Principal investigator of the VCU Work Incentive Planning and Assistance (WIPA)

National Training and Data Center, which provides training and support to 82 Social Security funded WIPA programs across the country.

Jeffrey B. Liebman *(Comment on Chapter 5),* Malcolm Wiener Professor of Public Policy, Kennedy School of Government, Harvard University—In his research, Dr. Liebman studies tax and budget policy, social insurance, poverty, and income inequality.

Gina Livermore *(Chapter 6),* Senior Fellow and Director, Center for Studying Disability Policy, Mathematica—Dr. Livermore's work focuses on improving the economic well-being of transition-age youth and working-age people with disabilities and includes numerous studies on the employment of Social Security disability beneficiaries.

Robert A. Moffitt *(Chapter 4),* Krieger-Eisenhower Professor of Economics, Johns Hopkins University and Bloomberg School of Public Health—Dr. Moffitt's research is on the economics of poverty and welfare programs and the economics of the labor market relating to the low-income population in the United States.

Austin Nichols (*Editor, Chapters 1, 3, Appendix),* Principal Associate, Abt Associates—Dr. Nichols is an economist affiliated with several national organizations. His work at Abt has focused on methodology and program evaluations in areas such as disability policy, return to work, employment and unemployment, housing, and education.

Sarah Prenovitz *(Appendix),* Associate, Abt Associates—Dr. Prenovitz is an applied micro-economist. Her research focuses on disability and education and has addressed the effectiveness and implementation of return-to-work, benefit navigation, special education, youth transition, and other programs for persons with disabilities. She has also studied the effects of the SSDI application process, school responses to incentives, the effects of educational interventions, and factors influencing career paths for PhDs.

Kathleen Romig *(Comment on Chapter 4),* Senior Policy Analyst, Center on Budget and Policy Priorities—Ms. Romig works on Social Security Disability Insurance, Supplemental Security Income, paid leave, and other budget issues.

Jesse Rothstein *(Comment on Chapter 2),* Chancellor's Professor of Public Policy and Economics; Faculty Director, California Policy Lab, University of California, Berkeley—Dr. Rothstein's research covers topics in education and labor market policy.

David Stapleton *(Comment on Chapter 9),* Independent Consultant, Tree House Economics—Retired from Mathematica in 2018, Dr. Stapleton continues to conduct disability policy research on a part-time basis. Much of his work has focused on how the Social Security Administration's disability programs affect the employment and income of people with disabilities, and the potential of policy reforms to improve income and employment.

Lucie Schmidt *(Comment on Chapter 6),* John J. Gibson Professor of Economics, Williams College; and Research Associate, National Bureau of Economic Research (NBER)—Dr. Schmidt is an empirical microeconomist working in the fields of labor and health economics and the economics of the family.

Jennifer Sheehy *(Comments on Chapters 5, 6),* Deputy Assistant Secretary, Office of Disability Employment Policy, US Department of Labor (DOL)—The mission of the Office is to develop policies that increase job opportunities for youth and adults with disabilities.

Jack Smalligan *(Comment on Chapter 2),* Senior Policy Fellow, Income and Benefits Policy Center, The Urban Institute—Mr. Smalligan analyzes the interactions across disability, retirement, and paid leave policy.

Vidya Sundar *(Chapter 8),* Associate Professor of Occupational Therapy, University of New Hampshire—Dr. Sundar's current research focuses on intervention programs for career development and sustainability for individuals with disabilities.

Till von Wachter *(Chapter 7),* Professor of Economics; Faculty Director, California Policy Lab; Director, Federal Statistical Research Data Center; and Associate Dean for Research, Social Science Division, University of California Los Angeles—Dr. von Wachter's research examines how labor market conditions and institutions affect the well-being of workers and their families.

Robert R. Weathers II *(Chapter 3, Appendix),* Chief Research Officer, Office of Retirement and Disability Policy, Social Security Administration (SSA)—Dr. Weathers's research focuses on the design and evaluation of SSA's random assignment demonstration projects.

David Wittenburg *(Chapter 6),* Disability Area Director, Mathematica—Dr. Wittenburg has worked on several research initiatives to improve adult outcomes of youth receiving Supplemental Security Income. He has worked on multi-site demonstration and research projects related to youth over the past 20 years.

Michelle Wood *(Chapter 9),* Principal Associate, Abt Associates—Her work has focused on leading project teams to conduct large-scale national program evaluations and applied social science research in areas such as disability policy, return to work, employment, housing, and homelessness.

Index

This volume draws lessons from more than 30 years of the Social Security Administration's demonstrations, which test new Social Security Disability Insurance and Supplemental Security Income policies, to improve outcomes and to highlight directions for future research.

This timely volume contains essays and insightful commentary from leading experts, not just about what we've learned about disability policy from past demonstration projects but also about how we can design even better demonstrations in the future to fill in the remaining gaps.
–Kathleen J. Mullen, PhD, RAND Corporation

This collection of papers is more than just a review of past findings… it provides new ideas for research demonstrations that move beyond just labor market outcomes and that could examine the broader issues of overall income security and economic well-being.
–David A. Weaver, PhD, former Associate Commissioner, SSA

This superb volume provides what has long been needed: comprehensive review and critical reflection regarding several decades of disability demonstrations. At once sobering and inspiring, it provides the indispensable foundation for the next generation of demonstrations.
–John Tambornino, PhD, Member, National Academy of Social Insurance