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

Overview

Over the past several decades, the Social Security Administration has tested many new policies and programs to improve work outcomes for Social Security Disability Insurance beneficiaries and Supplemental Security Income recipients. These demonstrations have covered most aspects of the programs and their populations. The demonstrations examined family supports, informational notices, changes to benefit rules, and a variety of employment services and program waivers.

A "State of the Science Meeting," sponsored by the Social Security Administration and held on June 15, 2021, commissioned papers and discussion by experts to review the findings and implications of those demonstrations.

A subsequent volume—*Lessons from SSA Demonstrations for Disability Policy and Future Research*—collects the papers and discussion from that meeting to synthesize lessons about which policies, programs, and other operational decisions could provide effective supports for disability beneficiaries and recipients who want to work. This PDF is a selection from that published volume. References from the full volume are provided.

Suggested Citations

Robert R. Weathers II and Austin Nichols. 2021. "Improving the Use of Demonstrations." In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Austin Nichols, Jeffrey Hemmeter, and Debra Goetz Engler, 85–134. Rockville, MD: Abt Press.

Jonah B. Gelbach. 2021. "Comment" (on Chapter 3: "Improving the Use of Demonstrations"). In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Nichols, Austin, Jeffrey Hemmeter, and Debra Goetz Engler, 127–130. Rockville, MD: Abt Press.

Elizabeth H. Curda. 2021. "Comment" (on Chapter 3: "Improving the Use of Demonstrations"). In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Nichols, Austin, Jeffrey Hemmeter, and Debra Goetz Engler, 131–133. Rockville, MD: Abt Press.

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.

Contributors

Robert R. Weathers II, 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.

Austin Nichols, 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.

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.

Jonah B. Gelbach, 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.

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.

Elizabeth H. Curda, 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.

Volume 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 54hat 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.