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

Burt S. Barnow and David H. Greenberg. 2021. "Design of Social Security Administration Demonstration Evaluations." In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Austin Nichols, Jeffrey Hemmeter, and Debra Goetz Engler, 31–84. Rockville, MD: Abt Press.

Jesse Rothstein. 2021. "Comment" (on Chapter 2: "Design of Social Security Administration Demonstration Evaluations"). In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Nichols, Austin, Jeffrey Hemmeter, and Debra Goetz Engler, 77–80. Rockville, MD: Abt Press.

Jack Smalligan. 2021. "Comment" (on Chapter 2: "Design of Social Security Administration Demonstration Evaluations"). In *Lessons from SSA Demonstrations for Disability Policy and Future Research*, edited by Nichols, Austin, Jeffrey Hemmeter, and Debra Goetz Engler, 81–83. Rockville, MD: Abt Press.

Chapter 2

Design of Social Security Administration Demonstration Evaluations

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

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

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

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

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

The 16 evaluations we reviewed are listed in Exhibit 2.1.

Exhibit 2.1. Reviewed SSA Evaluations

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

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

MAJOR EVALUATION DESIGN LESSONS FROM THE SSA EVALUATIONS

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

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

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

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

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

Non-Experimental Designs

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

Proof-of-Concept Studies

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

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

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

Non-Experimental Impact Studies

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

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

Use of Control Variables

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

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

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

Propensity Score Methods

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

Difference-in-Differences Analysis

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

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

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

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

Regression Discontinuity

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

Two Examples

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Experimental Designs

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

The Pros and Cons of Social Experimentation

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

Examples of SSA Experimental Evaluations

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

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

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

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

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

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

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

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

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

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

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

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

Sample Design Issues: Statistical Power and Minimum Detectable Effects

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

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

---

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

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

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

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

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

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

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

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

---

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

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

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

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

Population-Representativeness

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Outcome Measures

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Impact Estimation Issues

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

Data Sources

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

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

Statistical Approaches to Impact Estimation

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

---

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

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

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

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

Treatment of Missing Data and Missing Observations

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

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

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

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

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

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

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

Addressing the Multiple Hypothesis Testing Issue

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

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

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

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

---

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

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

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

Pooling across Sites

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

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

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

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

---

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

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

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

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

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

Length of Follow-Up

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

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

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

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

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

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

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

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

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

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

Efficacy versus Efficiency

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

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

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

The Role of Process Analysis

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

Implementing and Communicating the Intervention

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

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

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

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

Participation in the Intervention

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

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

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

Fidelity to the Intervention

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

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

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

Role of Cost-Benefit Analysis[34]

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

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

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

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

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

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

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

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

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

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

AREAS FOR FURTHER EXPLORATION

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

Alternative Experimental Designs

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

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

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

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

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

Clustered Designs

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

Stepped-Wedge Designs with a Staggered Rollout

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

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

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

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

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

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

Adaptive Designs

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

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

Factorial Designs

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

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

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

Other Experimental Designs

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

Seldom-Estimated Impacts

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

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

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

General Equilibrium Effects

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

Displacement Effects

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

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

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

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

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

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

Fiscal Substitution Effects

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

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

Equilibrium Wage Effects

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

Entry Effects

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

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

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

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

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

Program Component Effects

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

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

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

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

Site Representativeness

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

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

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

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

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

CONCLUSIONS ABOUT EVALUATION DESIGN LESSONS

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

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

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

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

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

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

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

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

Contributors

Burt S. Barnow, Amsterdam Professor of Public Service and Economics, The George Washington University—Dr. Barnow teaches a doctoral program seminar on public finance and human capital. His research focuses on evaluations of workforce programs and other social programs.

David H. Greenberg, Professor Emeritus of Economics, University of Maryland, Baltimore County—Dr. Greenberg is a labor economist and cost-benefit analyst. Much of his research focuses on the evaluation of government programs.

Chapter 2

Comment

Jesse Rothstein
University of California, Berkeley

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

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

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

WHAT TO EVALUATE?

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

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

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

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

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

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

---

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

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

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

STUDY IMPLEMENTATION AND POLICY

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

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

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

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

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

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

Jesse Rothstein, Chancellor's Professor of Public Policy and Economics; Faculty Director, California Policy Lab, University of California, Berkeley—Dr. Rothstein's research covers topics in education and labor market policy.

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

Chapter 2

Comment

Jack Smalligan
The Urban Institute

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jack Smalligan, Senior Policy Fellow, Income and Benefits Policy Center, The Urban Institute—Mr. Smalligan analyzes the interactions across disability, retirement, and paid leave policy.

---

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

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 58hat 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.