Doing Process Analysis Better

Michael Wiseman
George Washington Institute of Public Policy
WisemanM@GWU.edu

30 September 2012

Current Draft: 14 August 2013

Note: This report was prepared under contract to the U.S. Social Security Administration Office of Program Development and Research under contract HHSP23320095635WC/HHSP233337001T. This analysis is wholly the author's; it is not intended nor should it be interpreted as reflecting judgment or policy of the Social Security Administration.

For ease of reading, this draft is set up with hyperlinks between references in the text and the reference list for the paper as a whole; clicking on the year in (reference, year) takes the reader to the full citation.

GW Institute of Public Policy
THE GEORGE WASHINGTON UNIVERSITY

Contents

Doing Process Analysis Better

Summary

The draft *Demonstration Project Guidebook* prepared by the Office of Policy Development and Research in the Social Security Administration's Office of Policy defines a demonstration project as a "study that provides evidence of the feasibility and effectiveness of a new approach or practice." Demonstration plans are expected to include "impact evaluation, cost/benefit study, and process evaluation." This report to the Social Security Administration (SSA) reviews the current draft of the *Guidebook*, concentrating on what has traditionally been included, and what is generally needed, in process evaluation. Process evaluation is an application of process analysis, the investigation of how an activity is conducted.

This summary follows the outline of the main report but includes no references. More detail and references for matters reported in the summary may be found by going to the corresponding sections in the main report. The headings in the summary are hyperlinked to facilitate such investigation.

The Model

The diagram at right depicts elements of a policy demonstration; this helps situate "process." The "new approach or practice" is summarized by the program's design and intended management. The changes the demonstration tests typically have as focus certain target groups of people. Demonstrations are motivated by interest in a policy that is characterized in the model by both a design and a delivery plan. The intent is to change what happens to the target group. Everything occurs within a particular economic and social context that conditions outcomes.

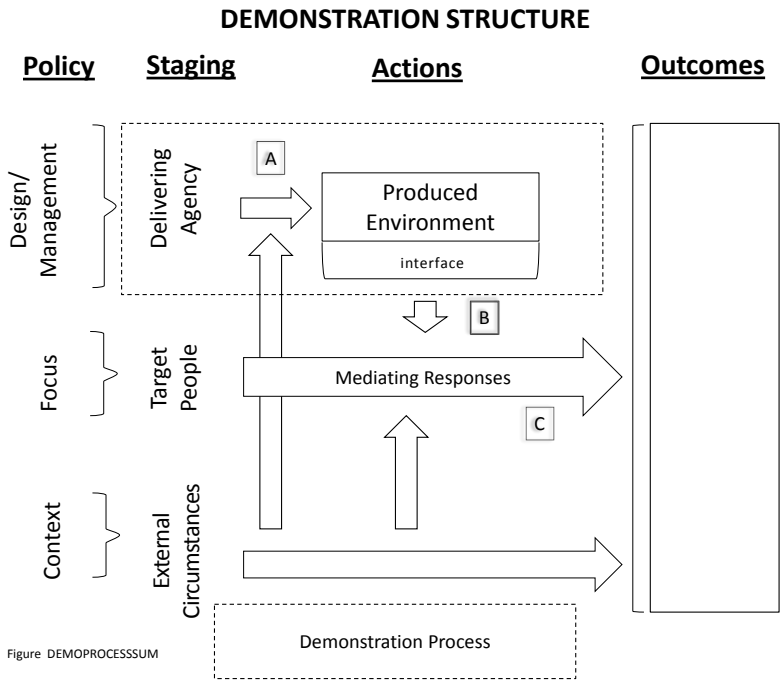

Figure DEMOPROCESSSUM

Four features of the model are emphasized. The first, identified by box A in the diagram, is the management process that connects the intentions of policy with the consequences as delivered to the target group. The second is that what counts for outcomes is what happens at the interface between the agency and the targets—box B. As depicted here, demonstrations change the environment of the target group; impact evaluation considers the consequences of this change in interface for outcomes that are of policy concern. Third, it is the response of the target group to the change in environment (as experienced at the interface) that is of interest. Generally the

theory behind innovations—often called the "logic model"—implies both intended intermediate and ultimate outcomes. The SSA's "Ticket to Work" initiative, for example, involves offering disability benefit recipients vouchers for employment training. Using the vouchers is a "mediating" response of the target recipients; employment is the ultimate objective. These mediating responses are identified in the diagram at box C. Finally, evaluating the impact of the demonstration requires prediction of outcomes in the absence of the demonstration—i.e., development of a credible counterfactual.

It is important to note in this context that a credible counterfactual is no easy concept. Measuring what the demonstration provides against what the theory of the intervention specified as the treatment (treatment fidelity) is important in interpreting outcomes. But in every impact evaluation there should be a second diagram that illustrates the process that defines the basis for estimating what would have happened to target group members had they not received what the demonstration provides. Thus, measuring the difference between the treatment and control experience also needs study of the control experience that matches study of the interface treatment the target group members receive. If this is not done, we cannot be confident that a similar innovation introduced elsewhere in time or place would have a similar impact. This diminishes the contribution of the findings to the social policy evidence base and to future multi-evaluation meta-analysis.

Process Analysis in Recent SSA Demonstrations

Three recent demonstrations exhibit aspects of the use of process analysis in SSA policy studies.

The Colorado WINS Youth Transition Demonstration. The Colorado experiment is one of six Youth Transition Demonstrations (YTDs) fielded "to improve understanding of how to help youth with disabilities reach their full economic potential." The YTD program model specifies that the six projects included in the evaluation be job-focused and provide employment services (emphasizing paid competitive employment), benefits counseling, links to services available in the community, and other assistance to youth with disabilities and their families. Participating youth are allowed to retain more of their disability benefits and health insurance while they work for pay. In Colorado, services were delivered by teams operating from the state's One-Stop Workforce Centers (now called American Job Centers).

The first-year Colorado YTD report is particularly useful, because it addresses all the features of the demonstration model shown in Figure 1 illustrates detection of a fidelity problem: The initial implementation in Colorado was inconsistent with the theory of the intervention. In working with treatment group youths, the service teams did not begin with employment issues, but rather first sought identification of more general goals and then moved to assembling services appropriate to these ends. This fidelity failure was detected by contractors and eventually remedied, an adjustment that underscores the importance of monitoring to successful demonstration implementation.

While treatment fidelity may have eventually been restored, the Colorado YTD evaluation is vague on services available to the control group. In consequence, the dimensions of the actual difference between treatment and controls are unclear, and the utility of the results for meta-

analysis is compromised. That said, it should be noted that the Colorado experience is probably not representative of the YTD effort as a whole.

The Mental Health Treatment Study (MHTS). The motivating hypothesis for the MHTS experiment was that a combination of access to a specific model of supported employment and "systematic medical management" services would enable and facilitate return to work by recipients of Social Security Disability Insurance (SSDI) benefits who have schizophrenia or an affective disorder. The template for integration of medical management and employment and employment support is called the Individual Placement and Support-Supported Employment (IPS-SE) model. The evidence base for IPS impact is considerable, and integration with Supported Employment was a natural next step in model development. SSA interest in IPS was encouraged in part by the growing importance of mental issues as qualifying disabilities for both SSDI and Supplemental Security Income (SSI). Fielded between 2006 and 2010, the MHTS experiment involved over 2,200 SSDI beneficiaries in 23 sites.

The MHTS evaluation illustrates the importance of the way in which the intervention is characterized. The study uses consistency of site-to-site utilization of services as a measure of uniform delivery of the IPS SE opportunity. But this confuses an intermediate output (service take-up), with the treatment (the opportunity). While this misstep does not compromise the assessment of the internal validity of the project's estimates of the impact of the opportunity for use of the IPS/SE combination on employment outcomes, it complicates the interpretation of efforts to analyze the effects of cross-site variation in program set-up and the fidelity of implementation. This in turn may weaken the utility of the results. As is true for the YTD study, the MHTS process component and meta-analytic utility of the study is weakened as well by failure to collect data on service environment for persons assigned control status.

While these shortcomings are important, the overall quality of the MHTS analysis is high, and the multi-site project setup provides possible opportunity for identification of the effects of treatment and context variation on program impact.

The Benefit Offset National Demonstration (BOND). Initiated in 2011, BOND is a 10-site national demonstration featuring enhanced financial incentive for return to work by SSDI beneficiaries combined with enhanced employment services management. The financial incentive is to allow beneficiaries who begin to earn amounts in excess of the "substantial gainful activity" (SGA) eligibility standard to retain eligibility for a considerably longer period of time than currently practiced. Moreover, instead of losing benefits entirely when earnings exceed SGA, beneficiaries in the treatment group have their benefits reduced by only $1 for every $2 of earnings in excess of SGA. The demonstration includes experimentation with "Enhanced Work Incentives Counseling" (EWIC) both in conjunction with the financial incentive and instead of it.

BOND is operating on a far larger scale than the MHTS or the YTD study. As conceived in the evaluation plan the BOND process analysis presents problems in implementation, focus, and policy use.

The implementation issues arise in the conflict between creating an SSA-like administrative environment for the BOND treatment group and staying out of the way of normal SSA

operations. Ideally BOND would deliver the combinations of employment incentives and service support in ways believed to replicate what would occur were such elements to be incorporated in regular social security operations. However, for various reasons SSA has difficulty delivering experimental treatments such as BOND in the context of common SSA office procedures. To avoid disruption of mainline functions, virtually the entire BOND operation is delivered literally outside SSA offices, by contractors. This poses a particular challenge for identifying how fidelity is assessed and how the interface experience is compared between experimental and control groups.

The Congressional mandate for evaluating the financial incentive called for fielding the demonstration at sufficient sites "to adequately evaluate the appropriateness of national implementation of such a program." The SSA chose to interpret this requirement to mean that the demonstration should be designed to produce results that would have external validity in the particular sense of supporting inference about what would have happened had the BOND been introduced nationwide to all recipients at the time of the experiment. The consequence is a muddying of the connection between the focus of the demonstration and general questions of SSDI policy. Moreover, it will be difficult to link BOND results to other, related projects in the formulation of evidence-based policy.

As was true for the MHTS, the BOND evaluation plan calls for using an outcome as a measure of fidelity of treatment delivery. It this case the outcome is the frequency of contact between employment counselors and members of the treatment sub-group eligible for EWIC. But in principle control group members, like most SSDI recipients outside the demonstration, already have access to the Work Incentives Counseling (WIC) program. The details on how the EWIC interface will differ from WIC are generally nebulous, with the consequence being that it is unclear how the practical difference between EWIC and WIC will be assessed. In general, the plan calls for gauging the treatment-control distinction through site visits and contacts with key informants. While there is some suggestion of feedback to on-going management from impressions gained from site visits, the process analysis is seen by the evaluators principally as providing *ex post* insight into the possible reasons for site-to-site variation in impacts.

The conclusion for BOND is that the effort to produce results that would be in a particular sense "nationally representative" appears to have compromised other goals of the project, most notably the ability to assess the actual change in environment created by the demonstration relative to either current SSA operating procedure as practiced or intended. The consequence is a muddying of the connection between the focus of the demonstration and general questions of SSDI policy. Moreover, it will be difficult to link BOND results to other, related projects in the formulation of evidence-based policy.

Looking across these three evaluations yields the implication that the strategy for process analysis might have been improved had the process analysis perspective discussed in this report been applied.

Process Analysis in Other Evaluation Guides

It is increasingly common for public agencies to develop manuals for project evaluation. The intended audience varies. Some, like the SSA's *Demonstration Project Guidebook,* are aimed at

personnel charged with overseeing agency-sponsored policy experiments. Others address organizations responsible for demonstration conduct. A third group addresses policymakers who need to understand evaluation theory and strategy. Virtually all discuss process analysis in one form or another. This report reviews process analysis discussions in three examples of guides currently in use.

Governmental Accounting Office (GAO). The GAO's 2012 manual is titled *Designing Evaluations*. The report is targeted both at the agency's own auditors/evaluators and at other agencies charged by Congress through the Government Performance and Results Act (GPRA) and other legislation with assessing program performance and impacts. Thus in many instances the GAO evaluates the evaluations of other agencies, and *Designing* in a sense provides the template against which such efforts are audited. Given the auditing emphasis of GAO activities, considerable weight is placed on evaluation *ex post* and the necessity of using a variety of methodologies to identify a counterfactual against which program impact can be measured.

The United Kingdom Treasury. Two documents set out Treasury's template for policy development. The first, the 2006 *Green Book* presents the government's general framework for project, program, and policy evaluation. The second, the 2011 *Magenta Book*, is the guidance specific to best evaluation practice for government departments. The importance of building evaluation into policy initiatives from the beginning is highlighted. The key *Magenta Book* problem lies in treating process and impact evaluations as if the concept of counterfactual arises only in the context of impact assessment and then principally in relation to outcomes, not inputs.

The World Bank. Like *Designing* and the *Magenta Book*, the World Bank's 2011 *Impact Evaluation in Practice* is in part intended to make the case for evaluation, but as the title indicates, the focus is almost exclusively on impact assessment, which the authors see as part of a broader agenda of evidence-based policy making. The Bank pays little attention to process analysis, adopting a perspective similar to that of the *Magenta Book*. Process is one thing; impact is something else. In this view, process evaluations are needed principally to track program implementation and as aids to interpreting the results of impact evaluation.

All three of the sample guidance documents discussed offer valuable insights into evaluation procedures. All three have shortcomings with respect the assessing the impact of an innovation on the environment of the target population—what is, from the perspective of my analysis, the essential objective of process analysis. The consequence is that evaluations following this type of guidance may be deficient for assessing replicability. They also may fail to produce the information on process impact that is important for accumulating knowledge for evidence-based policy.

Use of Process Results

Three examples of process analysis results illustrate the benefits of systematic attention to process. A 2003 study of techniques of process assessment for social services strategies for reducing drug abuse among prison inmates illustrates the procedures for systematic study of the interface between demonstrations and target groups and errors introduced by reliance on stakeholder interviews for information on actual program implementation. A 2012 meta-analysis of 15 years of welfare-to-work demonstrations indicates that impact evaluations of such

experiments reflect in part changes in what happens to members of the evaluation control groups, a theme consistent with my analysis. A second meta-analysis of welfare-to-work experiments, published in 2003, provides the most thorough integration of process with impact available in the social science literature. The study includes survey-based measures of key aspects of the project-target group interface. The results reveal substantial variation across experimental sites in the character of interventions and link this variation to significant variation in demonstration impact on earnings. This said, the study fails to achieve full comparability between data on the environments created by the demonstrations and the same-site environments of controls.

Process Analysis and Evidence-Based Policy

"Evidence-based policy" is widely endorsed, both by the executive and legislative branches of government and by outside organizations, notably the Coalition for Evidence-Based Policy (CEBP) and the (federal) Corporation for National and Community Service (CNCS). Both CEBP and CNCS have developed criteria for evaluating evidence, with greatest weight attached to inference based on "well-conducted" trials. But from my perspective, both the CEBP and CNCS criteria for qualifying evidence are puzzling. There is no component of the checklist of endorsed standards that relates to the description of process achievement or the experience of controls against which the treatment is measured.

The failure of the CEBP/ CNCS standards to include adequate consideration of process data has not gone unnoticed. The W.T. Grant Foundation has been particularly interested in improving understanding of the factors that influence the impact of innovations, and the implications of these moderating influences for predicting the consequences of "evidence-based" policy in education. The Foundation is currently promoting efforts to pay more attention to process in education-related demonstrations, but the principles carry over to SSA endeavors. Accumulating more experimental evidence—if combined with better information on treatment, control, and context—could, the Foundation argues, allow policy makers move beyond questions " 'what works' to learn to learn why and under what conditions programs are effective."

Conclusions and the Check List

What are the essential elements of process analysis? Logically, the conclusion is a list, applicable *ex ante* by the SSA staff members charged with evaluating "basic demonstration project development and design issues." The same list should be relevant to assessment *ex post* of demonstration results. I conclude from my review of selected process analyses of SSA and other demonstrations is that there are ten things to consider: (1) target, (2) treatment, (3) circumstance, (4) perception, (5) measurement, (6) production, (7) trajectory, (8) location, (9) choices, and (10) connection.

My report concludes by translating these 10 considerations into 10 steps to be taken in either planning or reviewing process evaluation. These are:

Step one: Review the logic. What is the target group environment the demonstration intends to produce, and how is that change expected to affect the outcomes of primary interest?

Step two: Distinguish processes. Separate the activities associated with establishing the counterfactual—that is, developing a prediction of what would have happened to treatment group

members in the absence of the innovation—from activities associated with delivering the treatment.

Step three: Identify the changes that count. Specify what change in the environment of the target group the demonstration treatment is expected to accomplish.

Step four: Find the measure. Develop a measurement plan for the environment the demonstration produces and the corresponding elements of the experience of the controls.

Step five: Anchor fidelity. Define the "fidelity counterfactual," i.e., what the intervention is supposed to deliver.

Step six: Consider connections. This ancillary activity should occur throughout demonstration planning; it considers the contribution of the experiment at hand to the policy evidence base and adjustments that might improve this connection.

Step seven: Think about trajectory. How does the analysis plan identify changes in treatment and control experience over the life of the demonstration?

Step eight: Specify the minimum. What are minimally acceptable accomplishments for the demonstration, and how can this achievement be ensured?

Step nine: Consider tradeoffs. Rank additional achievements beyond the minimum on the basis of rough benefit/cost assessment and develop an "ambition expansion path" through these potential achievements.

Step ten: Check with stakeholders. Throughout the planning phase confer with agency managers, those responsible for impact assessment, the benefit/cost accountant, and the sponsoring agency on the content and utility of the process analysis plan. In *ex post* assessment, identify process analysis shortcomings from the same stakeholders' perspectives.

Doing Process Analysis Better

Michael Wiseman

Introduction

The draft *Demonstration Project Guidebook* prepared by the Office of Policy Development and Research in the Social Security Administration's Office of Policy defines a demonstration project as a "study that provides evidence of the feasibility and effectiveness of a new approach or practice" (SSA 2010, 8). Such studies generally use social science methods to evaluate the consequences of some innovation in the way the agency goes about its business. The classic triumvirate of program evaluation is process, impact, and benefit/cost analysis. This report to the Social Security Administration (SSA) reviews the current draft of the *Guidebook*, concentrating on what has traditionally been included, and what is generally needed, in process evaluation. In doing so, my intent is to answer the question: What are the essential elements of process analysis?

The *Guidebook* instructs staff members who are preparing to champion a demonstration to answer a set of interrogatories both when evaluating a proposed project's "feasibility and effectiveness," and when addressing "basic demonstration project development and design issues" (SSA 2010, 18 and 19). The pertinent questions include reference to the classic triumvirate, albeit in an odd order:

> "Will the demonstration include an impact evaluation, benefit/cost study, and process evaluation? If any of these components are missing, how can we ensure that we will be in a position to answer all relevant questions concerning the effects and utility of the policy intervention that we are testing?" (SSA 2010, 50)

This, of course, begs the question of just what the essential elements of impact evaluations, benefit/cost studies, and process evaluations are. Answers are readily found in the literature and are certainly generally understood within the Social Security Administration (SSA). All three components of evaluation are detailed in the classic textbook *Evaluation: A Systematic Approach* by Rossi, Lipsey, and Freeman (2004). Nevertheless, the approach to evaluation by these authors is much different from that in, say, Michael Quinn Patton's widely used *Utilization-Focused Evaluation* (2008).

The variation in definition and approach across authorities is particularly notable in discussions of process evaluation, in part because process evaluation is included in what is more generally termed process analysis (i.e., the investigation of how an activity is conducted). Defined this way, process analysis covers a lot of ground. Thus, SSA interests are served by using the contributions of many authorities to sort process analysis out and to identify those essential elements. Such sorting and identifying is the object of this report.

Here's a roadmap for what follows: Section 1 develops a demonstration model and situates process analysis within it. Section 2 provides elaboration on the components of process analysis, the role of process evaluation within process analysis, and the interests process analysis serves. Section 3 reviews examples of process analysis in recent SSA demonstration evaluations.

Draft: 14 August 2013

Section 4 looks at the discussion of process analysis in evaluation guidelines produced by other agencies. Section 5 presents examples of uses of process analysis results, and section 6 links work at SSA to other efforts to promote "evidence-based policy." Sections 7 and 8 conclude with a vision of the role of SSA in promoting useful process analysis and a process analysis checklist for demonstration champions and critics alike.

1. How to Think About Demonstrations

Before introducing process, I review a common model of a demonstration or policy intervention. This is old hat, to be sure, but because of variations in component definitions, it is important to establish the definitions I will use in my later discussion.

The Demonstration Core

Demonstrations have three core elements: An action, a focus, and an intended outcome. The *action* is the "new approach or practice" to which the *Demonstration Guidebook* refers. The *focus* is who or what is to be affected by the action. The *outcome* is the consequence of the action. The demonstrations of interest to the SSA *Guidebook* have as focus certain people and as outcomes something about their behavior or circumstance. Figure 1 is a first demonstration depiction. Time runs from left to right, from staging to outcomes.

Figure 1: Simplified Demonstration Structure

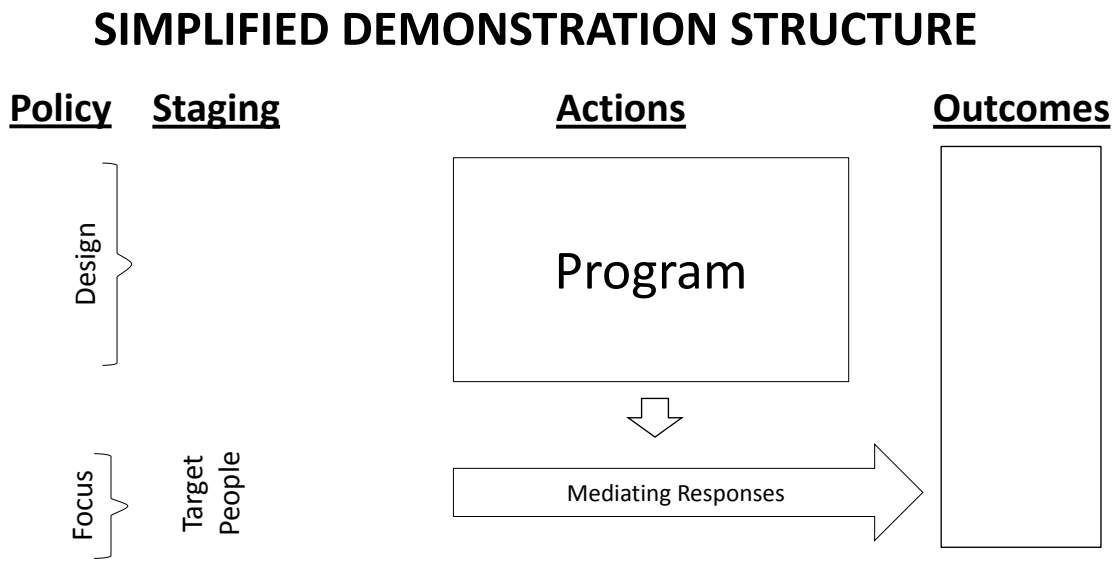

Figure SIMPLEDEM

As is done in the diagram, it is useful to draw a distinction between responses of the target population that occur as part of the action process, called *mediating* responses, and the outcomes of primary interest, the ultimate focus of the demonstration. It is common in the literature to call members of the target group "clients," especially when participation in the demonstration is voluntary. Demonstration planning typically begins with a story, called the program theory, that connects proposed actions with intended outcomes. Rossi, Lipsey, and Freeman call this "the set of assumptions about the relationships between the strategy and tactics the program has adopted and the social benefits it is expected to produce" (2004, 93). A demonstration's program theory is also referred to as the "logic model." A first task in identifying and addressing the "basic demonstration project development and design issues" cited at the beginning of my report is to work through what the underlying intervention theory—the logic model—is.

Consider as an example the SSA Ticket-to-Work program. (Detail and references are presented in the box.) The "ticket" is a voucher that recipients of Social Security Disability Insurance (SSDI) or Supplemental Security Income (SSI) benefits can use to obtain employment and other services they need to return to work or increase earnings. Services are delivered by state vocational education or other agencies that have been certified as eligible providers by SSA. The intent of the system is to encourage recipient efforts to prepare for and attain employment. Once a provider accepts a voucher, a service plan is initiated, and providers are reimbursed after attainment of employment goals the plan establishes. In the language of Figure 1, the targets are SSI and SSDI recipients, the policy action is the Ticket program, and the outcome of interest is employment. A crucial *mediating response* occurs when a recipient (in the language of the program) "assigns" his or her ticket to a provider (i.e., an "Employment Network" or EN). Note that in time (from left to right in the picture), the mediating responses precede the outcome of interest, employment.

In practice there may be no mediating response at all. Indeed the program may not affect the link between whatever the beginning states for the target people are and the outcome of interest. Tickets may not be assigned. Services may never be rendered, at least by the EN route. Nevertheless, some members of the target group will likely locate and utilize needed services, and some, with or without services, will become employed. The challenge is to identify the net effect of the Ticket intervention.

Detecting the difference between outcomes with and without an intervention like Ticket is the province of impact evaluation. Impact evaluation requires a counterfactual, an estimate of what would have happened to persons in the target group in the absence of the Ticket environment. Net effect or impact of the demonstration is gauged by comparison of outcomes with the innovation (the "treatment") to estimated outcomes without (the "counterfactual" or control). The net cost of the intervention is assessed by comparing costs of the treatment with cost estimates of what would have occurred anyway (the counterfactual). Combining the impact estimates with net cost provides an efficiency, or benefit/cost estimate. Easier said than done, to be sure! (Text continues on page 6.)

The Ticket to Work

The "Ticket to Work and Self-Sufficiency Program" was created by the *Ticket to Work and Work Incentives Improvement Act of 1999* (Public Law 106-170, TWWIIA) to expand healthcare and employment preparation services to individuals with disabilities to enable them to reduce their dependence on the income support provided by the Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) programs.

The voucher is the headline element of the initiative. In 2002, eligible SSDI beneficiaries (i.e., disabled workers, disabled widow(er)s, disabled surviving divorced spouses, and disabled adult children) and SSI disability recipients aged 18 through 64, began receiving a "ticket." This ticket could be used if the disability beneficiary wanted to obtain vocational rehabilitation, employment, or other support services from an approved Employment Network (EN) or State Vocational Rehabilitation (VR) agency of their choice. Participation in the Ticket to Work program is totally voluntary for beneficiaries.

National phase-in of the Ticket Program took place over a three-year period, and was fully operational by January 1, 2004. At the end of program rollout in September 2004, SSA had mailed Tickets to more than 11 million disability beneficiaries. SSA also modified the rules for the SSDI and SSI programs in order to give beneficiaries more incentives to participate in the ticket program. In addition to establishing the Ticket program, the Ticket to Work legislation increased access to Medicaid and Medicare services, provided for facilitated return to benefit for recipients who leave receipt for work but are subsequently forced to leave employment because of medical condition, and curtailed administrative emphasis on medical condition review (MCR) for recipients who begin training or employment. States were given funding to provide community-based work incentives planning and assistance (WIPA) to disabled beneficiaries to facilitate making informed choices about employment. These changes were expected to enhance the incentives for beneficiaries to join the program.

TWWIIA required SSA to recruit ENs, i.e., qualified private or public entities, to enter into an agreement to provide rehabilitation, employment or support services to eligible SSDI and SSI disability recipients. State VR agencies are allowed to participate as ENs.

An EN has the right to accept or reject a ticket based on their assessments of the needs of the individual and their ability to help that person. SSDI and SSI beneficiaries who decide to participate in the Program may take their ticket to any EN, or state VR agency, of their choice. The ticket provides evidence of SSA's agreement to pay ENs, in accordance with the regulations governing payment under the Ticket to Work program, for services rendered to the beneficiary. The ticket holder and the EN (or state VR agency) discuss employment options, goals, and a plan of services to attain these goals. If the beneficiary and the EN or VR agency agree to work together, they develop and sign an employment plan.

The Ticket program was set up on a "pay for performance" basis. ENs are paid when a beneficiaries achieve certain outcomes based on their employment. ENs cannot request or receive compensation for specific services from ticket holders. State VR agencies get paid on a case-by-case basis. They have the option of being paid under the existing state VR agency traditional reimbursement program or as an EN under the EN payment system.

While beneficiary use of the Ticket has grown over time, it is still low compared to the number of beneficiaries surveyed who expressed an interest in working. Between March 2004 and December 2004, the ticket participation rate rose from 1.1 percent of eligible beneficiaries to 1.4 percent in the states that first received tickets. However, Social Security disability beneficiary survey data suggest a much higher demand for employment and employment-related services. Twenty-six percent of disability beneficiaries surveyed indicated they saw themselves working for pay in the next five years, while 15 percent said they saw themselves earning enough to stop receiving benefits. Beneficiaries in the two later phase-in periods appear to be using the ticket at about the same rate as those who received tickets in the first phase-in period.

During the same time, only about a third of ENs had accepted any tickets, and many appeared to be losing interest in the program. ENs thought the Ticket administrative burden to be excessive, and beneficiaries took longer than expected to start work and have earnings at the level that resulted in a payment to the EN. Moreover, there was no incentive for beneficiaries who become gainfully employed to provide the EN with proof of their earnings, and this compounded provider problems.

Based on SSA evaluations and experience, several legislative changes were made in 2004 and 2008 to provide more incentives to beneficiaries, ENs, and employers. The 2008 regulations expanded the choices available to disability beneficiaries who want to enter or re-enter the workforce, encouraged more organizations to become ENs by revising their payment system to generate positive returns earlier, more often, and at higher rates, and promoted more partnering and coordination of services between organizations. A major Ticket outreach campaign was also initiated.

The emphasis of Ticket-to-Work evaluation has been on assessment of the mediating responses of ticket recipients and potential service providers. No impact assessment has occurred. However, the TWWIA also provided SSA with authority to experiment with altering the financial incentives for employment by introducing an earnings "disregard" in benefit administration for SSDI recipients. This "benefit offset" policy is the treatment in a national demonstration initiated (after a smaller-scale pilot) in early 2011. Process analysis in this Benefit Offset National Demonstration (BOND) is described in detail in the text.

More detail on Ticket background and operation is available on the SSA website (http://www.ssa.gov/disabilityresearch/twe_reports.htm).

---

*This summary is based in part on materials provided by M J Pencarski of the Social Security Administration. Ms. Pencarski's assistance is acknowledged with much appreciation.

Nuances

Before getting to process analysis, it is useful to add three nuances and augment Figure 1 to show them.

The first nuance is recognition that circumstances external to the demonstration may affect the mediating responses of the target group, the productivity of the innovation for the target group, and the outcomes directly. To continue the Ticket example, the local success of the voucher effort may be affected by the unemployment rate, both because impressions of likelihood of job-finding may influence whether or not target group members assign their vouchers and because the outcomes of Ticket services, once initiated, do indeed depend on the state of the labor market.

The second nuance is that some characteristics of the target groups may affect the productivity of the intervention, and others may not. Characteristics of the target groups as well as features of the environment that do affect the impact of the intervention are called *moderating*. For example, the impact of introducing something like the Ticket is surely influenced by factors such as the nature of a recipient's disability and his or her previous work experience. If so, these are moderators of program impact. Presumptions about moderating influences are typically the basis for division of the target population into subgroups for purposes of impact analysis. A program's logic model often includes assumptions or leads to presumptions about moderators.

Finally, passage of time produces many changes, some of which are relevant to a policy analysis, others not. Life goes on, sometimes without regard or response to the putterings of policy.

The upshot of these additions is an elaboration of the first model to include the influence of context, the distinction between moderating and other features of the target groups as well as the environment in which an initiative is launched, and a distinction between outcomes that are the focus of policy and outcomes that are not. These elaborations lead to Figure 2.

Figure 2: Simplified Demonstration Structure, Expanded

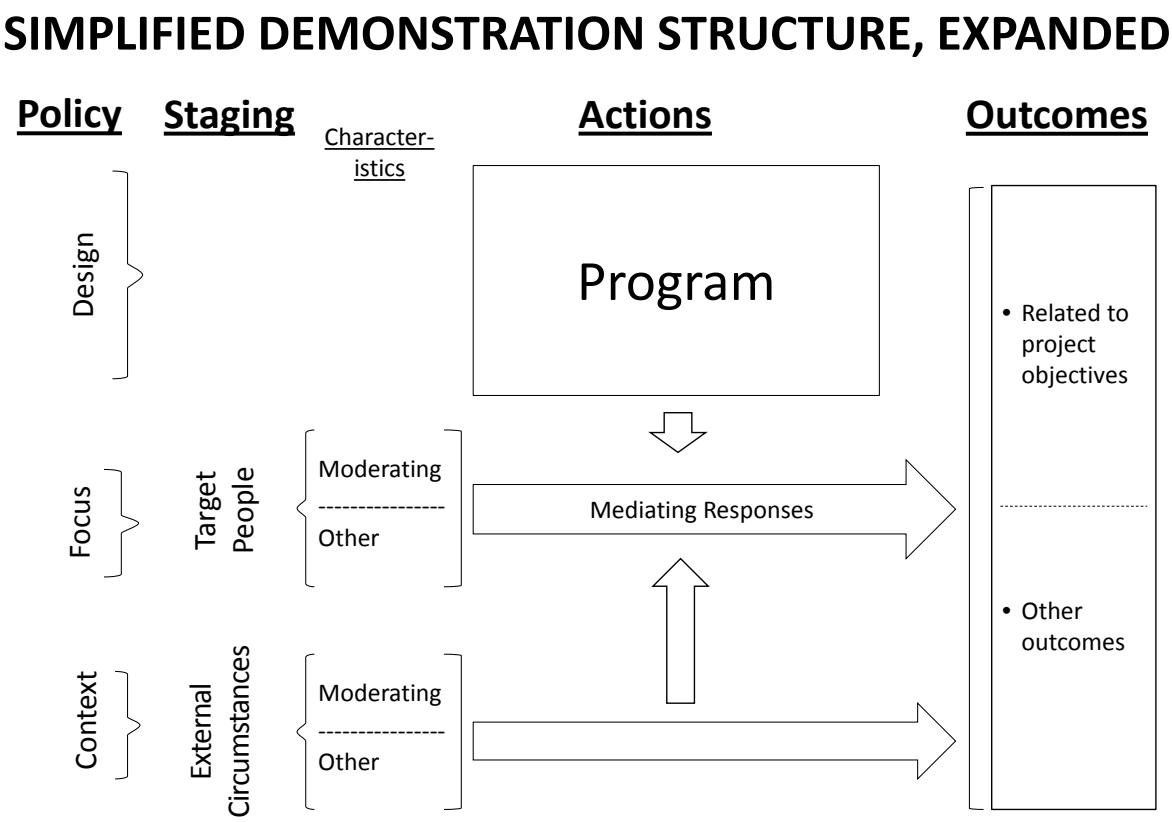

Figure SIMPLEDEMEX

2. Demonstration Process

The last step before getting to process *evaluation* is to return to process *analysis* and to distinguish two processes that are commonly the subject of attention. In the context of a demonstration, process analysis investigates how things are done. As used here, "demonstration process" refers to what it is an intervention delivers to targets, how this delivery occurs, and the potential consequences of what is achieved for project impacts. But in an evaluation context, demonstrations typically include an ancillary process set up to establish or identify a counterfactual. The primary interest here is in the process that creates what the program intends for participants, *not* how the impact of the program is evaluated, although the intervention production and evaluation processes cannot be kept entirely separate. The nature and consequences of their interrelationship will be taken up later.

The Interface-as-Outcome Approach

Here is the nub of my argument: The focus of demonstration process analysis as modeled here is the environment that introduction of "a new approach or practice" manufactures for the target clients. More specifically, the concern of process analysis is what a demonstration brings about at the *interface* between program and client—what the targets know and experience. As outlined

in the Box, the experience Ticket to Work intends to create is a set of options for obtaining information and supports plus a change (relative to earlier policy) in the consequences for clients of terminating benefit receipt as the result of employment. I use the active verb "create" to highlight that the environments envisioned by programs are the products of organizations and the responsibility of management.

This environment-as-outcome approach to process analysis brings us to the third (and last) demonstration feature to incorporate in the demonstration model. This is the production relationship that leads to the created experience. Thus, Figure 3 adds recognition that programs must be delivered, redefines "program" as "produced environment," includes both program design and planned management in policy, allows characteristics of the delivering agency to affect what appears at the interface between program and people, and mimics the subdivision of outcomes for clients by recognizing that the produced environment as experienced at the interface with clients always includes some features that are not the focus of, nor believed relevant to, process policy. One task of process analysis is to identify just what features do count.

Figure 3: Demonstration Structure, with Process

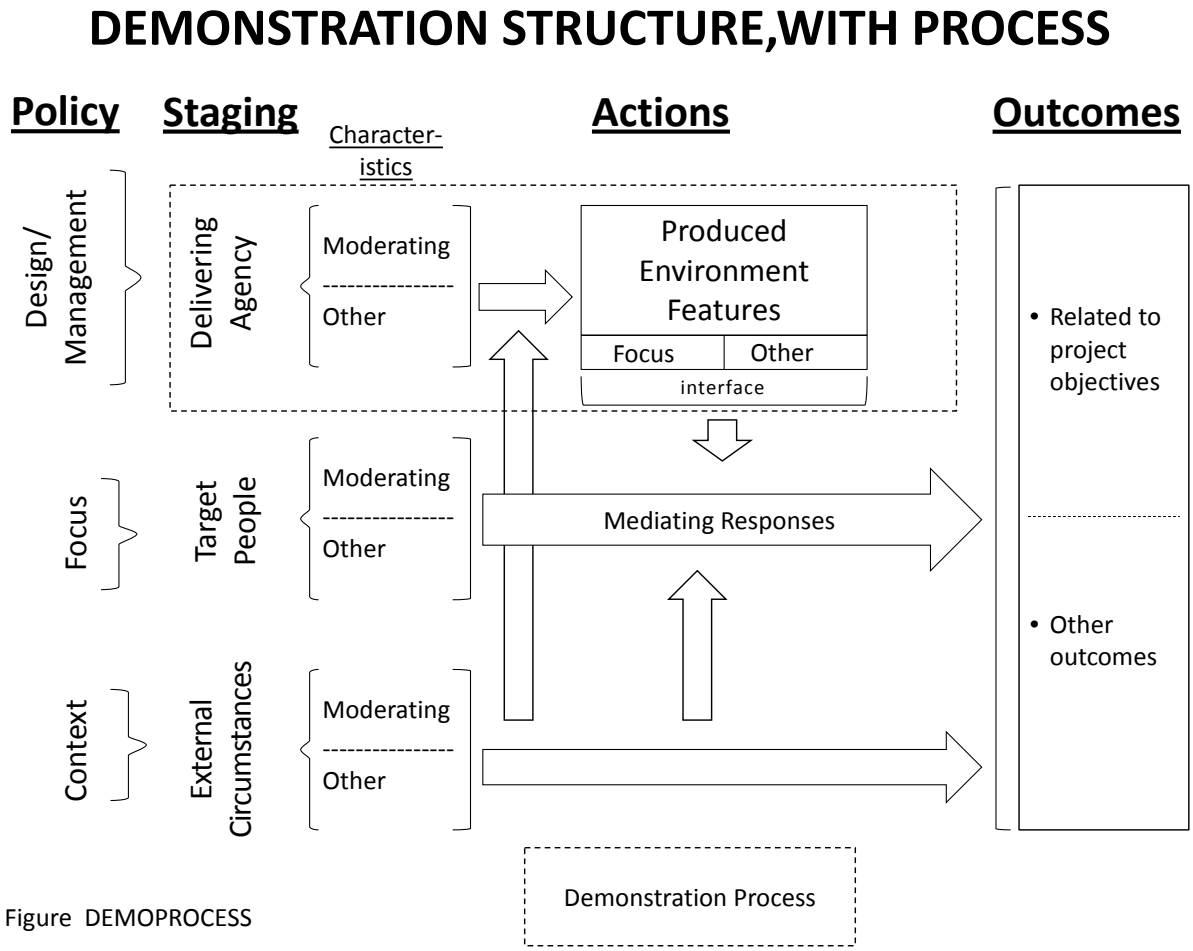

Figure DEMOPROCESS

Policies always include both intention (design) and delivery (management). It is useful to think of the origin of the experimental policy's design, focus, and management as the "home office"—for this discussion nominally the SSA headquarters in the Altmeyer Building in Baltimore, Maryland. But it is a long way from Baltimore to what happens to the targets of policy in, say, Tulsa. Every demonstration includes a staging institution with responsibility for delivering the interface the demonstration intends. The experience of this staging institution can be an important source of information about the feasibility of general implementation of the "new approach or practice."

In sum, process analysis studies what goes on in the Demonstration Box marked by the dotted line in Figure 3. The picture presumes that variations across delivering agencies—sites, perhaps—can moderate what the policy accomplishes. The diagram includes the possibility that the external environment influences the mediating steps in program implementation. One implication is that if a demonstration is carried out in multiple places, environments created may differ despite a common design intention. To return again to the Ticket example, it seems likely, and indeed evidence indicates, that even in context of a common overall design, the "feel" of Ticket-to-Work locally depends on the configuration of available service providers, the ENs.

Process Analysis and Process Evaluation

Process analysis has three components: theory, assessment, and evaluation. *Process theory* is analogous to the program theory or logic model of program impacts; but in this case the logic describes what might be called the "production function" for the demonstration, the management theory that connects the inputs assembled by the SSA and its agents to the interface between program and target recipients. *Process assessment* involves measurement of what the demonstration causes to happen for the target groups. And last, *process evaluation* compares what is produced at the interface to a standard or counterfactual.

Two types of counterfactual arise in process evaluation. One is a depiction of the intent of design, the environment the intervention is expected to produce. The Ticket to Work program is intended to offer SSI and SSDI beneficiaries choices among service providers and to provide the necessary information to support informed decisions. Comparison of accomplishment to program intent is typically called *fidelity* or management evaluation.

The alternative process evaluation counterfactual is another service configuration, possibly the set of services available before the intervention or generally in its absence. In a classical "randomized control trial" (RCT) the comparison environment is that experienced by the group exempted "by luck of the draw" from the innovation. The comparison of treatment to control environment is called an *impact* process evaluation. Other things equal—a presumption easiest to justify with random assignment—it is the impact on process that is responsible for whatever impact on outcomes a program accomplishes. To slip over for the moment to program evaluation strategy, interface impact is what is commonly called the innovation treatment—the difference between what happens to clients in the new program compared to what happens in a reference situation, the control. In practice, lack of fidelity is a suspect when demonstrations fail to achieve expected impacts on outcomes. But it is also possible that a demonstration fails to achieve significant interface impact because what is done for the target group differs little from the experience of controls.

Both management (fidelity) and interface impact evaluations require abstraction and symmetry. Abstraction is the identification of key features. Symmetry simply means comparisons require that these key program features be measured on both sides of the evaluation. For *management*, evaluation program theory tells us what the features should be; evaluation requires that we measure what they are. For *interface*, impact symmetry means measuring key features as experienced by both clients exposed to the innovation and clients in reference, i.e., control, circumstances.

To reiterate a bit, Figure 3 highlights an important aspect of SSA-type demonstrations. Most can be thought of as delivering two outcomes. One, the change in interface between program and target client, is the product of management. The other, the consequences for the target groups, is the product of the interactions among what management accomplishes, the characteristics of the target group, and the external environment in which the mediating responses of participants occur. Process theory focuses on the causal presumption linking management tactics to the client interface. Process assessment looks at interface content. Process evaluation considers the success of management in producing the environment called for by program theory. The measurement of and criteria for measurement success depend on the use to which the analysis is to be put.

The Purposes of Process Analysis: Four Perspectives

Figure 3 is complicated. What's important in demonstration planning? It depends on perspective. Four perspectives are relevant: management, internal validity, external validity, and building the evidence base. Differences among these perspectives plus intrinsic difficulties in observing process raise issues of costs and trade-offs. In this section the perspectives are considered first, then the trade-offs.

The *management* perspective is based on product delivery. What does the theory of the innovation require at client interface? How can these requirements be met across a variety of locations? This is the province of operations engineering. Managers look back from the interface to the components and mediating steps in the translation from the intent at Altmeyer to the interface in Tulsa. Managers are concerned with how the production process and delivered interface are moderated by the environment of the demonstration. The object of some demonstrations essentially ends with the management perspective: The intention is to determine whether it is feasible to deliver a program. If assessment of net effect on client outcomes is the object, however, the management perspective must expand to include production of the control or other reference intervention as well.

The *internal validity* perspective is concerned with interpretation of a demonstration's outcome as caused by the created interface. Internal validity depends on the credibility of the experience of the control as basis for forecasting what would have happened for the treatment group if presented with the control experience. Internal validity requires a reliable distinction between the environment created for the treatment and control groups and credible evidence that no systematic differences in unobserved characteristics between members of these two groups exist. Internal validity is thus a function of demonstration implementation.

Internal validity does not depend on successful production of what the innovation model requires. Instead, it requires measuring the differences between members of the treatment and control groups in terms of features identified as important. Measurement of this difference has as an important byproduct what might be called "descriptive validity"; the information provides basis for accurate summary of innovation accomplishment. There is, as a result, common interest from both management and internal validity perspectives in measuring what happens at the program interface for both treatment and control groups.

The *external validity* of estimates of demonstration impacts concerns the reliability of forecasts of the results of introducing a similar innovation in a location other than that in which the demonstration is initially carried out. External validity may depend on many things, including sensitivity of outcomes to variation in the control environment, variations in intervention as actually delivered, and the moderating effect of the peculiar circumstances of the demonstration location. Thus the *external validity* perspective begins with internal validity but moves beyond, to both local and more general application.

Local (sometimes called proximal) validity means the results of the demonstration apply to the location in which it was conducted. Local validity depends on the degree of difference, if any, between the experiences of controls in the demonstration and what is routinely characteristic of the experience of similar people as the program is currently operated in the site. There are three major threats to proximal validity. One is that the demonstration reference program, the control, does not adequately represent normal program functioning. A second is that the environment created as treatment in the demonstration cannot be "scaled up" to application to all clients, usually because production is contingent on inputs that are not readily cloned. Moderating features of the environment pose the third threat; if these features change over time, the relevance for tomorrow of what's learned today may be diminished.

General (sometimes called distal) validity concerns use of demonstration outcomes to forecast the impact of implementation of a similar treatment in another location. The major threats to general validity are lack of internal validity, disparity between what happens to controls within the demonstration and operations elsewhere, and differences between external moderating factors in the demonstration site compared to circumstances at the locations of new application. General validity is enhanced when a demonstration is conducted across multiple sites in a way that allows the consequences of variation in moderating factors to be gauged.

The *evidence base perspective* is related to the multiple-sites advantage. The evidence base for policy grows with the stock of demonstration results that have apparent general validity. Such growth occurs when demonstrations are replicated in sufficient detail to contribute to identifying the consequences of moderating factors and/or variation in features of the treatment or control programs. Demonstrations oriented toward the evidence base emphasize in their set-up the legacy of experience and features of the treatment or control experience that are congruous with other policy experiments. Establishing such congruity requires that the stock of available policy experiments include data on treatment dimensions as client characteristics and implementation circumstances. Thus, building the evidence base requires looking in two directions—*backward* to what has already been done and *forward* to the prospect of merging what may be learned with future experiments. It is to this external benefit that Rossi, Lipsey, and Freeman are referring

when they write that "the evaluation field itself becomes a stakeholder in every evaluation" (2004, p. 328).

While use of the term varies, my use of the term "meta-analysis" refers to efforts to combine results from multiple studies to reduce uncertainty about policy efficacy. In a sense, the evaluation *field* itself can be a stakeholder in an evaluation only insofar as the evaluation results can contribute to meta-analysis and, in combination with other studies, strengthen the evidence base for policy interventions. The implication of the Rossi-Lipsey-Freeman remark is that prospects for and contribution to meta-analysis should be part of every demonstration design.

All this is complicated within demonstrations by the presence of 1) processes intrinsic to the innovation and 2) processes that are part of the evaluation effort. Figure 3 is about the innovation. What must be kept in mind is that the two other pictures I have touched upon in the discussion so far. One is the corresponding picture for what happens to the reference group, the controls. In some demonstrations this is practice as usual; in others it can be another alternative to practice as usual. Whatever it is, understanding and measuring the counterfactual/control experience and the difference between treatment and control is essential to interpretation of demonstration outcomes, and to the local and general external validity.

The second alternative picture concerns the implementation of the innovation, how the treatment and control groups are established and sustained over the life of the demonstration, and how data are collected specifically for the evaluation. These processes are transitory by nature, since they exist solely to support the evaluation. But they can threaten local validity not only by producing differences between the environment experienced by the treatment group in a demonstration and what would be expected should the innovation be incorporated in normal operation but also by affecting the controls in ways that compromise their representation of a relevant counterfactual.

The multiple perspectives on process analysis lead to differing information requirements. Management wants more information on the mediating steps that connect organization to the interface outcome. Internal validity concerns call for information on the process of creating a forecast for experience in the absence of treatment, the counterfactual. Local external validity needs a gauge of difference between control and practice in place. General external validity requires information on moderating factors in the implementation environment. And building the evidence base calls wherever possible for deploying the same measurement procedures followed in prior work. *Choices are required* because information collection is costly and because, at some point, information collection must itself compromise both the local and general validity of demonstration results.

There is a (possibly shopworn) physics parallel here. In his famous 1927 paper, Werner Heisenberg asserted a limit to the accuracy with which certain pairs of properties of nuclear particles could be simultaneously known—for example, position and momentum. The better the information collected on one, the less reliable ("determinate") the information on the other. Process analysis may pose similar issues. Cost aside, the more information that is collected on what goes on in demonstrations, the greater the uncertainty about the relation between the experience of the target and control groups to what might be achieved or experienced in actual, non-experimental implementation. Thinking through the tradeoffs across perspectives—between resources devoted to collection of process data collection and resources devoted to expanding

sample size or measurement of outcomes, and between the gains from instrumentation and the impediments that information collection can create for service production—is a crucial part of demonstration planning.

Summary

This section has established a framework for process analysis that will serve as a guide for looking at the conduct of process analysis in recent SSA demonstrations. It encapsulates six major points:

1. Start with the interface. The beginning point for demonstration development is a statement of what the demonstration is to provide its targets. This treatment is the object of the demonstration process and the input to impact assessment.

2. Theory is essential. Demonstrations come with two theories. One, the management or production theory, details how the interface is to be created. The other, the intervention theory, explains how the interface is expected to affect outcomes. The management theory draws attention to mediating steps between policy concept and interface delivery and moderating factors that may influence the nature of the interface the demonstration creates. The intervention theory identifies mediating connections between interface and outcome and moderating factors that may influence project outcomes given the interface treatment.

3. There are two processes. Demonstrations come with two processes. One, the demonstration process, creates the interface. The other, the evaluation process, establishes the counterfactual against which the impact of the demonstration is measured. How the evaluation is done may affect the character of the intervention in ways that diminish the external validity of results.

4. Tradeoffs exist. Demonstrations address different concerns, and these differences lead to differing data demands. In consequence, the choice of objectives for a demonstration will affect the character and depth of the process analysis.

5. Explicit definition of the control experience is important. The headline feature of demonstrations is the treatment. But the success of demonstrations from every perspective requires careful attention to the referent experience, or control.

6. Measuring the counterfactual is important. Measuring the control experience, including over time, is essential for external validity and building the evidence base.

These concepts can be fleshed out by looking at the way they appear in current SSA demonstrations, the subject of the next section.

3. Process Analysis in SSA Demonstrations

In this section I review the crucial elements of three recent SSA demonstrations from the perspective of the process model depicted in Figure 3. The intent is to judge the utility of the process analysis model by application to important program experiments.

Process Analysis and the Colorado WINS Youth Transition Demonstration

The first example of process analysis in action is taken from the Youth Transition Demonstration (YTD), a multi-site evaluation of interventions intended "to improve understanding of how to help youth with disabilities reach their full economic potential."[1] The YTD portfolio includes six sites spread across five states. The projects began operations between 2006 and 2008 and lasted for three to four years (evaluation is ongoing, to be completed in 2014). Three of the projects were evaluated using random assignment. The evaluation contractors were Mathematica (lead), MDRC, and TransCen, Inc. TransCen is a nonprofit organization that participates in projects and activities involving school-to-work transition, education systems change, and employment for people with disabilities. TransCen is the technical advisory organization for the YTD projects.

This discussion considers only one of the individual YTD efforts, Colorado's "Colorado Youth WINS" project. The general YTD design includes one of the most thoroughly considered process analyses I reviewed for this study (Rangarajan et al., 2009). The Colorado YTD first-year report illustrates all features of the demonstration model presented in Figure 1, but implementation of the model at this site is a good example of process gone awry. We know this because of the process analysis apparatus put in place by the evaluators.

Set-Up

Each of the YTD projects was administered by a different lead organization, and the character and previous experience of the lead organizations varied substantially. The Colorado project was administered by Colorado WIN Partners of the University of Colorado-Denver. While site operations varied, each program was based on an intervention model that featured certain common elements:

> The YTD program model, which is based on best practices in facilitating youth transition, specifies that the six projects participating in the evaluation provide employment services (emphasizing paid competitive employment), benefits counseling, links to services available in the community, and other assistance to youth with disabilities and their families. Additionally, participating youth are eligible for SSA waivers of certain benefit program rules, which allow them to retain more of their disability benefits and health insurance while they work for pay (p. xv).

The Colorado project was carried out in four counties. The target population was youth aged 14–25 who were identified as SSI recipients from SSA records. The overall sample size was 468 in the treatment and 387 in the control group. The interim Colorado report used here includes detailed process and impact analyses; the impact study refers to the status of participants 12 months after entering the study.

[1] Fraker et al. 2011, xv. Page references in this section are to this source unless otherwise designated.

The Treatment

The Youth WINS services were delivered by "I-Teams" ("I" is for Independence) based in employment service One-Stop Workforce Centers.[2] The I-Teams were made up of a "disability program navigator," a "benefits planner," and one or more career counselors. Using a list of candidates drawn from SSA beneficiary records, the standard volunteer recruitment ("bait and deny") approach was used: The project and the evaluation processes were described to potential candidates. Those who volunteered took a baseline survey, signed a consent form, and were then assigned at random to a treatment or control group. Members of the control group were allowed to "receive only those services available in their communities, independent of the YTD initiative" (p. xvi); they were denied the YTD treatment.

The treatment is described in a succinct and telling way in the Executive Summary:

> The I-Teams were responsible for enrolling treatment group members in Youth WINS services. Through an intensive effort from August 2006 through May 2008, they obtained written consent to participate in services for 401 youth, or 86 percent of the treatment group members. Following their enrollment in services, the I-Teams sought to engage youth in discussions on a broad range of topics related to the transition to adulthood. From these discussions, the teams identified short- and long-term goals for the youth and incorporated these goals into evolving person-centered plans. The plans specified the services the youth required to achieve the goals. The I-Teams then arranged for those services to be delivered, either through referrals to other service providers or directly by the I-Team members. Youth were eligible to receive services for 18 months. Services were terminated in fall 2009, and the project formally ended in January 2010. (xvii)

This description incorporates some notable features of the report, including the following:

> While recruitment of candidates for the project was done by Mathematica, it fell to the I-Teams to enroll persons selected for treatment. This recruitment took a surprisingly long time, and resulted in conflict between time use for recruitment and time use for Youth WINS services. (43)

The treatment model, based on a system previously developed in Colorado, is a combination of team-based casework and service brokering on behalf of the client (note the reference in the program description above to "arrangement" of services). The model does not begin with employment issues, but rather calls for first identification of goals and then assembling services appropriate to these ends.

Employment is not mentioned in the source synopsis. This points to a fundamental problem with fidelity of the Youth-WINS operation to the general YTD model. Initially, employment services were not emphasized in treatment operations. When this shortcoming was detected by the

[2] In June 2012 the Department of Labor relabeled the One-Stop Workforce Centers the "American Job Centers." This report will refer to One-Stop Centers for activities occurring prior to this change. See Employment & Training Administration (2012).

evaluation team, pressure was placed on the Colorado WIN Partners to alter their strategy. This effort was only partially successful.

There are two horizons here. One, cited in the text of the quote, is the 18-month horizon, measured from the point of enrollment. The other concerns eligibility for the altered SSI rules applicable to YTD participants. These include a substantial enhancement of the disregard of earnings and provision for continuation of benefits regardless of outcome of a continuing disability review (CDR) and the age-18 medical redetermination. These treatment elements continued in effect for four years or to age 22, whichever was later.

The Process Analysis

The first Colorado YTD report is particularly useful because it addresses all the features of the demonstration model developed for the process study and illustrates a number of key problems. There are lessons to be gained here, and the contractors have done a good job identifying them.

Here are the features:

Target Group. The target group is young people aged 14–25 who, at the time of administrative record collection, were receiving SSI benefits on the basis of their own disabilities and lived in—or at least received mail in—one of the four target counties.

Intended Treatment. The intended treatment, common to all YTD projects, called for provision of "employment services (emphasizing paid competitive employment), benefits counseling, links to services available in the community, and other assistance to youth with disabilities and their families" (xv). It is significant that *employment services* comes first.

Administering Agency. The Colorado YTD project had an odd administrative structure that was the legacy of an earlier Colorado initiative funded directly by SSA. The I-Teams were employees of Colorado WIN Partners of the University of Colorado-Denver, but they were physically located within the county "One-Stop" workforce centers. The teams were nominally managed by the One-Stop center directors.

Delivery. The I-Teams were the delivery mechanism, but beyond case management the production process involved brokering of services from other agencies, most of which were represented at the One-Stops.

Management Feedback. The project used a special services delivery management information system called Efforts-to-Outcomes (ETO). I-Team members recorded all interactions. The basic source of information for the process analysis is the ETO data plus a survey

Mediating Outcomes. Given the model, there are important intermediate, or mediating, outcomes. These include development of a "Person-Centered Plan" and actual receipt of employment-oriented services.

Objectives. The central objective of YTD is assisting youth with disabilities reach their "full economic potential." Generally this means to become employed.

Lessons

In a nutshell, no progress toward the long-term goal of increased employment was evident at the end of the first year of Youth-WINS evaluation. The process study conducted by Mathematica and MDRC suggests this lack of progress was likely the result of failure by the delivering agency, Colorado WIN Partners, to ensure that I-Teams adopted the YTD intervention model. More than finger-pointing is involved. The case developed in the first-year report is quite strong. Obviously, much delicate diplomacy is involved here, and the first report includes a vigorous response by the director of Colorado WIN partners (in Appendix E). From the perspective of the process analysis project, the most important part of the report appears at the end of Chapter 3 as "nine key implementation lessons and challenges" (p. 63). An astute reader can infer all the key problems with YTD implementation from these.

Three points pertinent to the process analysis model developed earlier in this report and illustrated by Figure 1 deserve attention:

The first is that the Youth WINS process evaluation underscores important differences between evaluation process analysis requirements and management process analysis requirements. The evaluation focus is on gaining information about the change in environment that is actually accomplished for program participants. The evaluation perspective has a time dimension, since recruitment and treatment is distributed over a significant period of time, and what it is that the intervention accomplishes may change over the life of the project. But evaluators generally don't need this information immediately. In contrast, management requires knowing what is delivered quickly, to support adjustments intended to secure and sustain fidelity of the intervention to the project model. The Colorado YTD feedback to management was delayed and complicated by the complex management organization.

The second is that management matters. One intention of the present process analysis study is to draw a distinction between process outcomes and project outcomes. In the context of evaluation, process analysis attempts to measure the treatment actually delivered. But the fidelity of this product depends on technical feasibility, agency organization, and agency incentives. There seems little question that the intended Colorado YTD intervention was technically feasible. What appear to have been missing were the incentives necessary to ensure that CWP delivered. "Incentives" do not refer necessarily to cash. A big incentive is provided by what good process analysis makes public about management achievement, or lack thereof.

Finally, the Colorado YTD report underscores the notion that in innovation planning, it is important to pay attention to the starting point. CWP had prior experience with a case-management/service brokering approach to services for youth with disabilities. This history provided a useful base of experience, but at the same time it carried the culture of services with goals other than employment. The base for the agency is, after all, a *medical school*. Clearly more thought should have been given the culture issue in planning for YTD implementation. The first-year report refers to this issue when listing the following as #3 under "key implementation lessons and challenges" (p. 63): "When scaling up programs that require a shift in focus to adhere to a conceptual framework, having strong management buy-in is an important factor in successful implementation."

Shifting focus is not what "scaling up" is about. Youth WINS was not a scaling up project in the sense the term is used elsewhere in the literature. The "shift in focus" essentially created a new intervention.

Issues

This review of the first-year Colorado Youth WINS report has turned up several issues. Here are three:

What Happened to the Transition? Given the attention paid the issue in most discussions of SSI for children, it is surprising that virtually no attention is given in the Colorado YPT report to the effect of the mandatory age-18 medical CDR on participant responses to recruitment or, once assigned to treatment, response. One of the "gifts" of the YTD waivers is assurance that participants will continue to receive benefits for four years or to age 22, whichever comes later. The utility of this gift depends on whether one has made it past the CDR. Youth were recruited for the Colorado YTD from both sides of this divide. What would we expect to be the consequence?

What is the Scope for Meta-Analysis? The multi-site YTD project includes a number of different approaches to employment-oriented services, but the target population is the same. It is not clear how data from Colorado will be combined with the data from other locations to develop more general estimates of impact.

What Happens to Controls? Perhaps the most glaring deficiency in the Colorado report is failure to provide information on the service environment for the controls. The process analysis includes two interesting vignettes describing the opportunities presented by Youth WINS changes to "Ian" and "Ashley," chosen "to profile youth who were active participants in the project" (p. 37, n. 50). What's missing is a depiction of what would have been available to Ian and Ashley's respective counterparts in the controls. As emphasized in the process analysis model, measuring the difference between the opportunities for treatment-Ian and Ashley and control-Ian and Ashley is essential to modeling net impact and to meta-analysis.

The control group in the Colorado YTD poses a number of problems. There is evidence that some controls were exposed to treatment services via participation of the I-Teams in the One-Stop context. Also, cutbacks in general disability services in Colorado are an important external development. Evidence within the Colorado report suggests the cutback may have differentially affected the treatment and control groups. If so, this development will be difficult to observe given the instruments in place for process analysis.

The Mental Health Treatment Study (MHTS)

The MHTS provides a second "process analysis in action" example. Again, I interpret the intervention from the perspective of the model demonstration structure presented in Figure 3. The source for this discussion is the project final report (Frey et al. 2011), and chapter-page numbers listed are to that source.

Set-Up

The motivating hypothesis for the MHTS experiment was that a combination of access to a specific model of "supported employment" (SE) and "systematic medical management" (SMM) services would enable and facilitate return to work by recipients SSDI benefits with schizophrenia or an affective disorder.

The template for SE delivery in MHTS was the Individual Placement and Support (IPS) model. IPS content is well defined as the result of numerous previous trials and documented fidelity assessment procedures. IPS SE encompassed a variety of features with emphasis on rapid movement into competitive employment through individualized job search and supported by both on-going, work-related vocational assessment and job development. Services made available in the SE treatment package continued beyond job-finding to include on-the-job support and on-going effort to assist in improving employment status. The SE services were combined with other behavioral health (OBH) and related services thought to contribute to employment acquisition and continuation, health, and quality of life.

Medication is an important part of modern mental illness treatment. SMM services in the demonstration used a Nurse Care Coordinator to coordinate and monitor drug treatments prescribed for participants, to promote preferred medication practice, and to monitor maintenance of drug regimens. In addition to SE and SMM, the treatment group received various other behavioral health and related services, a comprehensive insurance plan that covered most out-of-pocket medical expenses, and suspension of SSDI CDRs for three years. The CDR suspension was intended to remove the threat of loss of SSDI eligibility due to demonstration of capacity for "substantial gainful employment," thereby encouraging employment.

The MHTS was fielded by the SSA between November 2006 and July 2010 in 23 sites across the United States. Over 2,200 SSDI beneficiaries participated. As is the case for my two other process analysis examples, the impact of the intervention was evaluated using the volunteer participant strategy. Volunteers with the target primary diagnoses were recruited for the project from recipient rolls, informed of the nature of the project and the evaluation, ask to certify informed consent, and then allocated at random between the treatment and control groups. The primary source of information on outcomes is interviews done at quarterly intervals over the 24-month participation interval following baseline data and consent collection and random assignment. The analysis approach is again primarily intent-to-treat (ITT); effects are assessed by comparing outcomes for all treatment group members to all controls.

In collaboration with SSA, the MHTS evaluators identified four primary study outcomes and a variety of secondary outcomes. The focal primary outcome is employment; the others are physical health status, mental health status, and participant perceptions of quality of life. Employment impacts were substantial, as were positive consequences of treatment assignment for mental health and quality of life. After 24 months there was no significant difference in physical health between the treatment and control groups.

The MHTS evaluation reports distinguish between evaluation and treatment processes.

Evaluation Process

The evaluation process analysis has three major components: site selection, response to opportunity, and verification of random assignment.

Site selection. The 23 sites were selected to provide regional diversity (the Southwest is the only region not included), but evaluators were primarily concerned with identifying locations with a Community Mental Health Center (CMHC) capable of providing the mental health services and with a history of providing SE in a manner meeting IPS standards. Two sites operated without a Center connection were expected to provide comparable services through other institutional arrangements. It is easy to understand the strategic utility, indeed necessity, of seeking sites with adequate CMHCs, but this selection strategy undercuts any argument that the broad regional representation supports presumption of external validity of the demonstration results. The effect of site-to-site variation in external circumstances is generally left out of the analysis, except for inclusion of local population density as a determinant of participant mental health (4-48).

Recruitment occurred from the SSA Master Beneficiary Record (MBR) database based upon age, location, and recorded primary diagnoses of SSDI recipients. Approximately 62,000 candidates were identified, and 57,000 letters of invitation were mailed.[3] A wide range of outreach approaches were used as letter follow-ups; nevertheless one-quarter of beneficiaries could not be located using MBR address data and another 8,000 did not respond to phone calls. Ultimately 3,971 candidates attended a Research Information Group (RIG) meeting to hear about the study, and 2,238 were enrolled in the program. (Some RIG attendees and other respondents were found ineligible due to recent SE involvement or living under guardianship.) Efforts at letter follow-up were not consistent across sites, because recruiters were working to achieve site sample quotas and would cease pursuit of additional candidates once quotas were achieved.

Response to Opportunity. Since the "universe" from which recruitment was undertaken is well-defined, it was in principle possible to include within the analysis package a study of determinants of willingness to participate. However, since outreach efforts were not uniformly exerted across the initial candidates list, the MHTS evaluators elected to create three universe subgroups (3-4). The first, "potential enrollees" (PE), is made up of all beneficiaries contacted by voice minus those subsequently determined ineligible for the project. The PE group includes roughly 16,000 beneficiaries. The second, "possibly potential enrollees" (PPE), comprises 14,000 beneficiaries who (presumably) received the MHTS invitation letter but were not reached by phone, either because the maximum number of attempts was achieved, no calls were made, or calls were made but not recorded. The third, "not potential employees," consists of 32,000 beneficiaries not located, never spoken to, nor determined ineligible. The response-to-offer analysis is then conducted by studying the factors that differentiate those who agreed to participate in the experiment from, alternatively, the potential enrollees group or the combination of the PE and PPE groups (termed the CPPE group). There are no responders in the PPE group; all of the enrolled group is contained within PE. Thus response rates based on CPPE assume that every member of the PPE group is disinterested in MHTS.

[3] Numbers in this paragraph are taken from Figure 3-1, p. 3-4 in Frey et al. (2011)

Constructed in this way, both PE and PPE groups were divided into equal test and validation samples. A logistic regression model of the likelihood of MHTS participation was then experimentally constructed using the PE test sample. Independent variables for the model were drawn from the MBR, the only source covering the entire CPPE group. Once built, the model was then re-estimated for both the combined test sample and the PE and combined validation sample. The results are consistent across the PE and combined samples, but the estimated probabilities of participation are reduced by half, of course, by inclusion of the PPE in the denominator.

The validity of the response-to-offer analysis is problematic given uncertainty about MHTS information actually received by members of the PPE group, but the results reasonably support a number of inferences about the targeting of MHTS-type initiatives. The importance of targeting is underscored by the very low incidence of interest in the initiative regardless of the universe subset used as the base. The targeting implications are advanced with appropriate cautions in the project summaries. The quality of the results could have been enhanced, possibly significantly, by a different strategy for beneficiary pursuit and better record-keeping. However, given that RTO was not a major focus of the demonstration, small flaws on this feature of process may be of little import.[4]

Random Assignment Validation. Random assignment was validated by the usual comparison of characteristics of the treatment and control groups. In general there were no significant differences in composition between the two. Oddly given the point of the study, the most important difference appears in primary diagnosis. Approximately 68 percent of the treatment group had an affective disorder, compared to 72 percent of the control group (3-14). This difference is mirrored in the greater prevalence of schizophrenia within the treatment group.

Treatment Process

Treatment process analysis in the MHTS report is divided between implementation of the IPS SE and OBH and related services components and the SMM and NCC services.

IPS SE. The description of SE services investigation reveals what are, from the perspective of my analysis, significant issues. The description includes an assessment of several different aspects of implementation. Measurement at the site level assessed the extent to which the study sites implemented SE services as intended. That is, did the study sites develop and provide the kind of SE services that are consistent with the IPS model of demonstrated effectiveness? This consistency is what the evaluators call "program-level fidelity" (5-1). Measurement at the individual level assessed the extent to which treatment group participants received the array of SE assistance and OBH and related services expected in a program that is faithful to the IPS model. The relevant term for this level is "beneficiary-level fidelity." A third aspect of implementation assessed in the study, also at the beneficiary level, was the extent of active beneficiary involvement in the intervention offered by the study site. The relevant term for this is "engagement." (Frey et al 2011, 5-1)

[4] In a larger context, what may be more important is the apparent unreliability of address information in the MBR. Benefit checks presumably are handled through direct deposit. But how are CDRs managed without addresses?

Program level fidelity assessment is comparison of what is offered to a standard, in this case the IPS SE model. The emphasis is on what the operating agency in each site delivered as opportunity to participants and the consistency of that package with the template. The demonstration featured ambitious efforts both to train site operators in the nuances of the IPS SE model and to monitor implementation throughout the study.

It is in the interpretation of "beneficiary-level fidelity" that problems arise. The MHTS report seems to confuse an intermediate *outcome*—utilization of services—with uniform delivery of the IPS SE *opportunity*. Utilization of services is identified within the framework developed here as a mediating response (again, see Figure 1). From the perspective of this report, what the MHTS authors term "program-level" fidelity is a matter of having the necessary machinery for delivering IPS SE and SMM in a site. Beneficiary-level fidelity is the opportunity presented to the treatment group: Does management make the machinery deliver the environment for MHTS participants called for by the intervention model? Finally, given that, what is the response of participants? The intermediate or mediating response, engagement, is a function of the supply of IPS SE/ SMM opportunity provided by each site and the demand by participants for those services. "Supply" can include marketing.

The confusion between mediating and target outcomes does not threaten the reliability of the analysis of treatment/control comparisons, since this work is done solely on the basis of outcomes unadjusted for mediating response by the treatment group. The authors find no significant correlation between measurements of the program-level fidelity across sites and employment rates achieved by the treatment group, but this may simply reflect success in monitoring and ensuring model implementation at this level (5-25). But the confusion does affect analysis of variation in outcomes among treatment participants. In the MHTS report, participant engagement is measured by use of services, and participants were scored on the basis of frequency and breadth of use. The resulting score is incorporated as an independent variable in analysis of variation among treatment group members in employment-related outcome such as the likelihood of obtaining employment (4-23). This raises issues of simultaneity: It would seem reasonable that participants who are inclined to work and see opportunity for employment are more likely to take advantage of services available. Thus the estimated impacts of engagement (the mediating response) on the likelihood of obtaining employments, retaining a job as a "steady worker," or duration of job holding—all focal outcomes— may be spurious.

Anecdotal evidence reported in the analysis indicates that significant variation did occur across sites in the degree of integration of IPS and other behavioral health services, but integration was not measured, in part due to lack of "psychometrically-validated, comprehensive, multi-item scale using multiple data sources" (5-4). This is unfortunate, because integration is generally an operational matter, something that affects the nature of the MHTS opportunity as it confronted participants. Had a measure been developed, it could have been included in the models of determinants of variation in treatment group outcomes. Used this way, integration could be appropriately treated as an exogenously determined characteristic of the treatment group's experience.

The analysis strategy reflects an important ideological aspect of the IPS movement. According to the report, "The IPS model assigns responsibility for engagement to the IPS program (which includes integrated behavioral health, such as case management) rather than attributing lack of

engagement to a shortage of motivation on the part of IPS participants. In this sense, the IPS programs fell short in employing effective engagement strategies" (5-23). Thus the treatment interface needs to include producer commitment to pursuing participants to the point of achieving an engagement standard, something that is difficult to imagine how to do were IPS SE to be implemented generally. But note the report's reference to "IPS *programs"* (emphasis added). There appears to have been substantial variation across sites in pursuit of engagement. For example, 69 percent of MHTS enrollees received benefits counseling at some point during their MHTS experience (5-14). But this rate *varied from 9 to 98* percent. What could have been done would be to model the likelihood of obtaining this (or any other) IPS service based on participant characteristics at baseline *and site.* Once this model was estimated, the *predicted* probability of engagement could have been entered into the model of determinants of employment outcomes for the treatment group as an instrument for variations in the character of IPS SE as actually delivered at the interface. Note that this approach allows/expects some variation in engagement success on the basis of client characteristics, a possibility not acknowledged by IPS ideology.

NCC-SMM. As with IPS SE, the Nurse Care Coordinator/Systematic Medical Management dimension of treatment was assessed at two levels. Program-level fidelity was enhanced by the fact that the program itself provided an NCC at each site. It was complicated by arrangements for medication prescribers. The CMHCs that were in most sites the hosts for the program had prescribers in place. Since the template for SMM and the active role of the NCC were not typical of standard operating procedures, on-site prescribers had to operate on a dual track, using SMM for one group of patients, but not for another. The report notes that this "likely affected the degree to which they bought into and adopted the SMM program as their *modus operandi"* (6-3). Moreover, a significant proportion of the treatment group came into the experiment with an established relationship with a prescriber with no connection to the site's supporting clinical system. The evaluators expected such prescribers to be difficult to engage fully in SMM (and to relate as intended to the NCCs) "for both logistical and systemic reasons" (6-3).

The prescriber problem made beneficiary-level fidelity to the SMM component problematic. Here as in IPS-SE, the analysts focus on measures of intermediate outcomes, ratings of beneficiary engagement in SMM and prescriber engagement, both as reported by NCCs. There is considerable variation across sites in beneficiary SMM engagement, ranging between 75 and 100 percent for engagement in any of the reporting periods and between 25 and 87 percent for engagement over the entire span of project record-keeping (6-18). This variation is clearly associated with location of prescriber; off-site prescribers were far less likely to be "fully engaged" in SMM.

Given the evidence that off-site prescriber location led to reduced participation in SMM, at first consideration it might appear that variation across participants in initial use of on- versus off-site prescribers could serve as a useful instrument for variation in full access to SMM services. However, this is complicated because many participants did not come into the program with any established prescriber relationship, so the presence of one might well signal exceptional problems that would lead, with or without SMM, to differences in employment outcomes when compared to those without. And some participants switched from previous prescribers to those on-site. To their credit, the evaluators did not include a measure of individual SMM participation in their analysis of determinants of employment-related outcomes for the treatment

group.[5] Nevertheless, site differences in SMM participation appear to be an important matter for future investigation. If it can be determined that some part of the variation is due to factors beyond participant control, it may be possible to use such factors to aid in identifying SMM effects, something that is not achieved by the MHTS analysis as currently presented.

The Controls

Chapter 8 of the MHTS report is devoted to study of utilization of healthcare services by the treatment and control groups. With one important exception, what is measured (by quarterly surveys) is the incidence of psychiatric emergency: emergency room visits, hospital admissions, nights spent in hospital, crisis services. From the perspective of the model I use in this report, these are all outcomes. These outcomes are of interest both as a manifestation of impacts of the MHTS intervention on matters other than employment and because changes in utilization of such services brought about by the intervention affect overall benefit/cost assessment for the project. Comparison of service use between treatment and control groups indicates that in MHTS treatment significantly reduced hospitalization and use of crisis-oriented outpatient services.

The (partial) exception to the focus on outcomes is a measure of "other clinic or mental health provider visits." This measure, the MHTS report notes, is reasonably treated as a product of the design of the intervention: "The increases in insurance coverage for behavioral health and the provision of systematic medication management services presumably encouraged treatment group participants to make ongoing use of behavioral health services" (8-9). This encouragement had consequence: The treatment group averaged 52 visits over the two-year project horizon, compared to 35 for the controls (8-6); the size and statistical significance of the estimated difference grows with progressively more elaborate multivariate control (8-8).

However, the "other visits" variable may also tell us something about variation across sites in the opportunities the project provided to participants. Presumably some of the observed variation across sites in the frequency of other visits for participants with similar backgrounds reflects differences across sites in delivery of the access to the medical services that are an important element of the SE-SMM model. It might be possible to develop a measure of the probability of use of other services by site and to use this measure as an instrument for SMM/case management delivery fidelity, just as the prevalence of benefits counseling was suggested above as an instrument for measuring variations in the character of IPS SE. However, rather than exploiting inter-site variation in implementation as an opportunity for identifying SMM impacts, the MHTS report treats such variation simply as a source of error correlation across observations taken from the same site or as fixed effects to be addressed with indicator variables for site.

Assessment

The internal validity of the primary MHTS results appears unquestionable. This was an exceptionally ambitious experiment, and the criticisms voiced here need to considered in that light.

[5] SMM engagement is used as an independent variable in modeling variations in mental health status at study exit for members of the treatment group. A statistically significant positive association is reported.

The process analysis is problematic. On the one hand, the effort at achieving fidelity appears to have produced reliable evidence that what was available at sites corresponded to the prescription of the IPS model. On the other hand, the character of SMM appears to have varied because of both site factors and participant situation on program entry.

For both the IPS-SE and NCC-SMM sides of the analysis, the report confuses mediating outcomes—engagement—with treatment. Despite achieved program-level fidelity, it is not clear what the sources of variations in engagement across sites are. One of the most questionable aspects of the analysis is use of individual engagement rates as right-hand variables in models of the variation in employment outcomes across treatment group members.

Perhaps the most puzzling aspect of this demonstration is that, in the end, what was managed was verification, on a heretofore unobserved scale, of the efficacy of the IPS ES model. But the evidence for IPS-ES was substantial even before MHTS. What this demonstration intended was a test of the impact of adding SMM and professional coordination of care. Yet the execution failed to do this. The problem was not just that there was no treatment group with IPS-SE and no SMM. There apparently was no consideration of systematically varying the intensity of efforts to engage participants in both SE and SMM services in order to see if more effort had a payoff. If it did, it would have demonstrated the underlying effectiveness of more aggressive case management and medications management, even without the no-SMM treatment. As an alternative, consideration should be given to modeling variations in engagement in both IPS SE and SMM across sites, and using the predicted values of engagement as an instrument for variation in program interface in models of outcomes for individual participants.

A similar problem arises with process in the last of the three examples, the Benefit Offset National Demonstration.

The Benefit Offset National Demonstration (BOND)

BOND is a response to a Congressional mandate contained in the Ticket to Work and Work Incentives Improvement Act of 1999. In addition to setting up the national system of voucher-based training and supportive services provision for SSDI recipients described earlier (see box, p. 4), the *Ticket* legislation encouraged SSA to implement and evaluate incorporation of an enhanced financial incentive for return to work by SSDI beneficiaries. As implemented, the innovation includes experimentation with enhanced services management. After a pilot demonstration of the BOND concept in four states beginning in 2005 (Weathers and Hemmeter 2011), a 10-site national demonstration was initiated in early 2011. BOND is quite complex; I concentrate on the process analysis conducted for the services component.[6]

[6] Unless otherwise noted, the description that follows is drawn from two reports by BOND implementation and evaluation contractors, Abt Associates and Mathematica Policy Research. See Stapleton et al. *BOND Implementation and Evaluation: BOND Final Design Report* (2010, henceforth *Design Report*) and Bell et al. *BOND Implementation and Evaluation: Evaluation Plan* (2011, henceforth *Evaluation Report*). As would be expected, various features of the design and evaluation have been altered in response to implementation experience. My concern here is with the original plan.

I first summarize where the services component fits in, then outline the evaluators' objectives for the process analysis. Viewed from my perspective here, the BOND process analysis presents problems in implementation, focus, and utilization.

Work and SSDI

Understanding the policy in place is the beginning of appreciation of the BOND intervention and point of reference for evaluating BOND implementation and outcomes. I concentrate on SSDI. Slightly fewer than one in seven SSDI recipients qualify for additional benefits through SSI (SSA 2011, Table 66). Such "concurrent" recipients have low SSDI entitlement and few assets. SSI operates as a standard "negative tax" transfer program with the same impairment eligibility standard as SSDI. This complication is ignored here but discussed in detail in BOND planning documents (cf. Stapleton et al. 2010, p. 12).

Workers become eligible for SSDI benefits after establishing a required history of work and contributions to the SSDI Trust Fund. Claims are engendered by the onset of a disability. A disability is defined as a physical or mental impairment that renders a worker unable to engage in "substantial gainful activity" (SGA). The impairment must be "medically determinable" and expected to last for at least a year or to result in death. SGA is defined in terms of potential monthly earnings. In 2012 the SGA standard is earning net of impairment-related work expenses (IRWE) in excess of $1,010 per month. A higher standard applies to the blind.

Persons seeking SSDI benefits go first to local Social Security field offices. The field office confirms that the applicant is earning less than SGA and has a qualifying employment history. These hurdles passed, the case is referred to the state's Disability Determination Service (DDS) for disability determination. If benefits are "awarded," payments begin in the fifth month following the onset month for the disability. Because many awards are made only after appeal, the elapsed time between onset and first payment can be much longer than five months (although retroactive payments will be made as appropriate). The payment level is a function of previous earnings. The average payment in 2010 was $1,068 per month (SSA 2011, Table 2).

As the name indicates, SSDI is an insurance system, not a means-tested public assistance system. The logic of the insurance system is that a claim is made on the basis of an entitlement, and the award covers a fraction of earnings lost due to disability onset. The replacement continues as long as the disability continues. When awards are given, a "continuing disability review" (CDR) is scheduled on the basis of assessment of the likelihood and timing of recovery. These reviews are typically planned for three or seven years past the onset date, depending on the nature of the disability; SSA workloads often push the effective CDR date beyond this.

People do recover from disability or at least regain the capacity for SGA. The insurance-program logic implies that once capacity for SGA is regained or discovered, benefits should cease. However, even when employment can be found, returning to work presents considerable uncertainty and may be perceived by beneficiaries as risky, especially because of the connection between SSDI receipt and access to Medicare. Amendments to the Social Security Act in 1960 added to SSDI regulations provisions for graduated return to work. As currently operated, the law provides a 45-month period for SSDI beneficiaries to test their ability to work without losing all benefits. This interval has two parts. The first is the "Trial Work Period" (TWP), nine

months within any five-year period in which the beneficiary can work any amount, including earning beyond SGA, without effect on any SSDI-related benefit. Work that counts for TWP accumulation is determined by an indexed threshold monthly earnings total. In 2012 the TWP threshold was $720 per month. Once the TWP is completed, any month of earnings above SGA leads to termination of benefits, although beneficiaries are allowed the cessation and two additional "grace" months before benefits are withdrawn. Cessation of benefits and the grace period is followed by a 33-month "extended period of eligibility" (EPE) in which benefits are resumed for any month in which earnings fall below SGA. Medicare eligibility extends through the EPE and for an additional 54 months, as long as the disability endures.

Obviously, the system is complicated. There is virtually no evidence base for either the structure or parameters of the system. As configured, the return-to-work process is challenging to administer in a reliable way. Among other things managing the TWP, EPE, and Medicare extensions requires substantial bookkeeping and elaborate records organization. With or without rationale, the operation is difficult for beneficiaries to grasp. And even with the TWP provisions, any beneficiary who reaches the end of the TWP has much to lose by working beyond SGA. Indeed, if working beyond SGA signals recovery from the disability upon which original SSDI qualification is based, benefits (including Medicare) simply end after the three cessation and grace months.

Working at all, let alone working beyond SGA, is a rare occurrence for SSDI recipients. Liu and Stapleton (2011) report that among workers awarded SSDI benefits in 1996, only 6.5 percent experienced benefits suspension over the coming decade and less than 4 percent had benefits terminated after finding work. These low rates may simply reflect the severity of disablement required to pass the SGA criterion. But they may also reflect the behavioral consequences of both the income loss that earnings moderately in excess of SGA, if sustained through a TWP, produces and the tortured and uncertain path the system creates for those recipients who want to work.

Credibility may be part of the problem. The SSA devotes considerable effort at promoting work, among other things by producing and distributing red, white, and blue brochures on the pathway to work and features of the system that accommodate return to employment by beneficiaries who wish to do so. The 2012 SSA brochure *Working While Disabled—How We Can Help* (http://www.ssa.gov/pubs/10095.pdf) cites 102 Work Incentives Planning and Assistance (WIPA) projects across the country that provide Community Work Incentive Coordinators (CWICs) to help SSA beneficiaries make "informed choices about work." But given the low prevalence of the experience of return to work, it is unclear just how much beneficiaries know and how confident they are about the benefits and risks of moving to employment (Livermore 2003).

The complexity of the SSDI work incentive complicates modeling of the impact of SSDI structure on behavior. The standard abstraction for modeling treatment of earnings under SSDI is the single-period labor-leisure diagram that highlights the SGA "cliff" and predicts bunching of recipients at earnings levels just below SGA (cf. Weathers and Hemmeter 2011, p. 712).[7]

[7] Weathers and Hemmeter cite a U.S. General Accounting Office (2002) for "a description of this behavior," but fail to note that the GAO report provides no evidence that it is significant.

Anecdotal information plus common sense suggests that caution would be in order for an SSDI recipient contemplating return to work and that the income loss associated with sustained earnings above SGA is a disincentive. "Common sense" may also include recognition that sustained earning above SGA belies the notion that a recipient continues qualification for SSDI, regardless of the letter of current regulations. This inconsistency is a specter that haunts much of the discussion of incentives in disability policy.[8]

One of the more interesting developments with respect to incentives for employment is an option, confirmed in 2008, for Ticket-to-Work service providers, the ENs (see Box, p. 4) to make payments to beneficiaries for work-related expenses that are very broadly defined. If the claims are large enough, this amounts to conversion of the Ticket-to-Work payment into cash. At least one national EN (AAA TakeCharge) promotes this option. The beneficiary assigns his or her ticket to the EN, an employment plan is established, and as earnings and benefit objectives are attained, the beneficiary receives 75 percent of the outcome payment received by the EN. These payments can be large—as much as $19,278 in Work Support Payments over three years for a beneficiary who begins earning at SGA levels or above. I have found no economist's description of the SSDI SGA "cliff" that mentions this parachute or considers its incentive effects, yet currently AAA TakeCharge claims to have more Tickets assigned than any other EN in the country.

The BOND Innovation

BOND is an experiment with altering the treatment of earnings after completion of the Trial Work Period, cessation, and grace periods to gauge the effect of increasing the financial incentive for sustained employment at earnings levels above SGA. Congress specified the incentive in some detail:

> The Commissioner of Social Security shall conduct demonstration projects for the purpose of evaluating, through the collection of data, a program . . . under which [SSDI] benefits . . . are reduced by $1 for each $2 of the beneficiary's earnings that is above a level to be determined by the Commissioner. Such projects shall be conducted at a number of localities which the Commissioner shall determine is sufficient to adequately evaluate the appropriateness of national implementation of such a program. Such projects shall identify reductions in Federal expenditures that may result from the permanent implementation of such a program.[9]

Developing the required demonstration projects presented both architectural and strategic challenges. Determining the threshold level of application of the "1 for 2" benefit reduction (called an "offset") is an example of the architectural problem. Choice of a threshold level below SGA would reduce benefits for employed SSDI recipients with earnings above the threshold but less than SGA. It is doubtful that Congress intended any group of recipients to be disadvantaged by an experiment with alternatives. Likewise, the legislation says nothing about the endurance of the incentive. Demonstration of a "forever" offset would create a substantial

[8] Non-insurance disability assistance systems tend to emphasize the obligation of the capable to re-enter the labor force, but enforcing obligation presents other problems.
[9] Ticket to Work and Work Incentives Improvement Act of 1999, Public Law 106-170, 113 Stat. 1902 (1999).

and difficult to budget financial liability for the agency. On the other hand, a short-duration offset might not provide sufficient incentive for detectable effect.

SSA chose to set the threshold at SGA and the horizon at five years past conclusion of the TWP. Therefore an SSDI recipient in BOND earning more than SGA could retain full earnings disregard for the duration of TWP, cessation, and grace months and lose $1 in benefit for every $2 in earnings above SGA for 57 additional months. Moreover, while benefits continue to be paid monthly, accounting under BOND is shifted to an annual basis, so that in effect most beneficiaries will receive payments based on average monthly earnings. Payments are established annually based on the basis of expected earnings. Recovery of overpayments or restitution of underpayments occurs in the year following accumulation. Mid-year adjustments may be made in projected earnings in light of actual experience; such adjustments should reduce accumulating payments errors.

The strategic problems include ensuring that recipients eligible for the offset understand it. The BOND treatment is a substantial alteration in incentive and payments strategy. As anticipated and confirmed by results from the BOND pilot demonstration, communicating the new environment to beneficiaries is problematic (Weathers and Hemmeter 2011). In principle the WIPA Community Work Incentive Coordinators could, with training, shift their story for BOND participants from the old to the new system. BOND designers included an ancillary innovation of substantial enhancement of engagement and information called Enhanced Work Incentives Counseling (EWIC). EWIC counselors have smaller caseloads than Community Work Incentive Coordinators and are intended to have more time, more skill, and more resources to use in engaging recipients in efforts to work.

BOND is implemented in the coverage areas of 10 SSA Area Offices. These sites were selected to be, with appropriate weighting, nationally representative. The consistency of this strategy with the Congressional mandate "to adequately evaluate the appropriateness of national implementation of such a program" is not clear (more on this external validity issue later), but neither is the meaning of "appropriateness" in this context. In each site the benefit offset is evaluated in the two stages presented schematically in Figure 4. In the first, all SSDI recipients in the project areas are divided at random into three groups: a control group, a treatment group, and a solicitation group. The treatment group is informed of the BOND $1 for $2 alteration and given information on where to obtain additional WIPA-type counseling on how the benefit works. Since for practical purposes no SSDI recipient is made worse off by the offset, no "informed

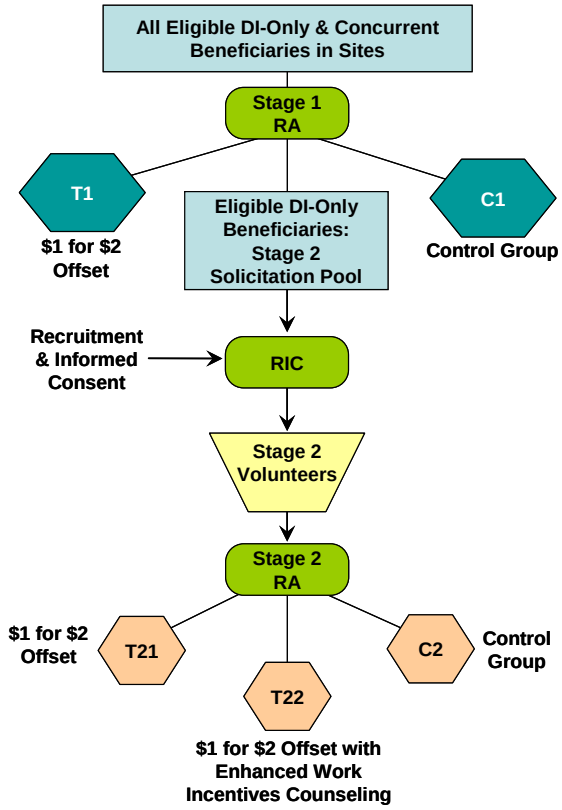

Figure 4: BOND Random Assignment and Sample Design

consent" is sought. Treatment group members are simply informed of their altered state and opportunities and directed for assistance to BOND-trained WIPA CWICs.

Members of the solicitation pool from Stage 1 are then recruited for a second stage that includes the opportunity for EWIC. Once the Stage 2 project has been described, the pool candidates are invited to participate. If they volunteer, they are asked to confirm informed consent and to complete a baseline survey. Again at random, volunteers are assigned to one of three groups. The control group continues with current practice, including WIPA incentives counseling. A second group receives the $1 for $2 offset and WIPA counseling modified to reflect the terms of the benefits computation procedure. The third group receives the offset with EWIC. The intention is to gauge the impact of the change in incentive alone as well as any additional gain from more aggressive counseling engagement.

BOND will last a long time. Participation is limited to beneficiaries who complete their TWP by September 2017. Those who happen to complete in this last month will have 60 months of participation opportunity, so it is possible that the last $1 for $2 benefit adjustment will occur for September 2022. The EWIC treatment for Stage 2 will terminate in September 2017.

BOND Processes

The principal outcomes of interest for BOND are impacts on recipient earnings and benefits. The logic model for the demonstration sees any effects on earnings and benefits as generated by the change in the interface between the system and beneficiaries that will be brought about by the demonstration. The basic change is the offset. The secondary change is introduction of enhanced services.

On the production side for the offset, the challenge is not only to produce the intended offset opportunity for both Stage 1 and Stage 2 treatment group members but also to ensure that the opportunity is understood to a degree and in ways consistent with what would likely be achieved in national implementation. The same is true for the EWIC program for the second-stage services. Some of the mediating outcomes—notably beginning Trial Work Period—occur prior to actual utilization of the offset. Any effect therefore depends upon what is perceived about the intervention by members of the treatment group. Perceptions are hard to measure.

The evaluation process presents two types of problems. The first is randomization and sample development. While administratively challenging, procedures for random assignment are well understood. There is some evidence that time as a recipient may have a mediating consequence for intervention effects, specifically that short-term SSDI recipients are more likely to respond to the incentive than are those with lengthy program experience. Accordingly, the BOND target population was divided between recent awardees and all others, with recent awardees oversampled. "Recent" means cases open 36 months or less at the point of sampling.

The more difficult problem of the evaluation process involves the controls. According to the evaluation plan, the intent of the demonstration is to support inference about the impact of BOND implementation nationwide. Nationwide at the time of BOND implementation, various administrative problems complicated the conduct of the Trial Work Period, and working recipients' actual status can be a matter of uncertainty. Is the reality of implementation at the

point of BOND implementation the appropriate point of reference, or should it be a normative control, faithful to the concept, if not the practice, of the SSDI return to work process as currently operated?

BOND Set-up

One very important question about the BOND treatments concerns the producing agency. The ideal is to replicate for the treatment groups the interface that would exist with full integration into SSA operations. In an experimental context this is problematic, especially when the treatment populations are small relative to the overall caseload. The strategy pursued for BOND was to cede responsibility for most operations and for the evaluation to a contractor-collaboration. As originally proposed, the production and evaluation processes were clearly demarcated:

> To ensure the objectivity of the evaluation, we divided the BOND team into two components. The implementation team is responsible for setting up and operating the demonstration, including enrollment of subjects, assisting SSA with administration of the offset, recruiting counseling organizations, training counselors, and overseeing the delivery of WIC and EWIC services to BOND clients. The evaluation team is responsible for random assignment of beneficiaries to the various BOND (i.e., treatment) and non-BOND (i.e., control) groups, data collection to support the evaluation, analysis of the data, and reporting the findings. Evaluation team staff will play no role in implementing the intervention, and the implementation team will have no input into the evaluation. This bifurcation assures that the BOND research findings reflect the independent, unbiased assessment of the evaluation team researchers (Stapleton et al. 2010, p. 8).

The process analysis component of the BOND evaluation plan conflates evaluation of the recruitment and random assignment processes with evaluation of interface achievement. The focus of the recruitment analysis is study of the determinants of take-up of the Stage 2 offer. The focus of the evaluation of interface achievement is on fidelity of actual BOND processes to the intervention model. These processes include the employment monitoring and payments system and the EWIC.

In general, the planned analysis of fidelity is based on a series of site visits by review teams, including a pre-implementation visit to all sites to establish, through key informant interviews, a baseline picture of service availability. The fidelity analysis extends to evaluation of control group experience and confirmation of the differential between benefits-related WIC counseling and the more general set of services provided through EWIC.

Much of the evaluation of fidelity will be based on ratings by site visit teams. The BOND plan emphasizes the distinction between EWIC and WIC by developing a series of quantitative indicators of fidelity. For example, for the "outreach and engagement" dimension of the demonstration, EWIC counselors are expected to be in contact with beneficiaries on at least a monthly basis, while WPA counselors provide only group sessions and seminars. The process plan calls for gauging this by documenting the frequency and character of contact between the system and clients, with a fidelity benchmark for EWIC of 88.5 percent of treatment group

members engaged in EWIC services in any month (Bell et al. 2011, p. 87). Considered in relation to the model developed for this analysis, this engagement rate is an outcome, not necessarily a measure of fidelity of the EWIC offering to the demonstration model. But like the discussion of service participation in the MHTS, variation in predicted rates of contact across sites and over time could be treated as an instrument for variations in supply.

The problem of confusion of achievements at the interface with outcomes appears at other places in the process evaluation plan. For example, the EWIC benchmark for "referrals and coordination with organizations that provide job placement" is that "60 percent of T22 [stage 2 EWIC treatment group] subjects who engage in EWIC services will have earnings at some time during the project" (Bell et al. 2011, p. 89). More significantly, while the measures proposed may provide useful information on variation across sites in the character of the interface created for the treatment groups, very little information is available on variation across sites for the controls.

Part of the problem may reflect failure to implement the distinction between implementation and evaluation. In fact there are three BOND activities: (1) delivering the innovations; (2) establishing the difference between innovation and counterfactual, and (3) evaluating the difference between innovation and counterfactual both for input and outcomes. Task (1) is a major problem in management, and while it is likely that "the process study team will use the same definitions, indicators, and benchmarks as the implementation team to assess the fidelity of WIC and EWIC services" (Bell et al. 2011, p.), it is also likely that production will require more.

While the common use of these indicators by the implementation and evaluation teams suggests some feedback from process outcome to on-going management, the process analysis is seen by the evaluators principally as providing *ex post* insight into the possible reasons for site-to-site variation in impacts. The final "Cross-Cutting Analyses" chapter of the *Evaluation Analysis Plan* lists seven process measures to be used in assessing "whether any patterns emerge in the relationship between the participation and the process findings" (Bell et al. 2011, p. 170). The measures include things like measures of processing time for the offset and whether or not sites "met work-focused interview benchmarks." However, a footnote declares that "formal statistical analysis of the relationship [between participation and process achievement] is precluded by the limited power provided by the sample when process findings are available for just 10 sites" (Bell et al. 2011, note 119). Quite apart from whether achievement of benchmarks is an appropriate approach to obtaining measures of process, this "power" statement ignores the possibility that the character of process, both for treatment and controls, may vary over time. The invariance with respect to time of process ranking across sites is an artifact of the evaluation plan data collection strategy; it may bear little relation to what occurs on the ground, for either control or treatment groups. To be sure, data collection costs might well have ruled out anything more ambitious, but the alternatives appear never to have been investigated.

BOND Critique

The BOND process analysis plan poses several important questions.

External validity. If executed according to plan, differences in outcomes between the treatment and control groups created by the demonstration can be attributed to the differences the experiment creates for the SSDI interface for the various groups. Given the competence of the implementation team, the internal validity of attributing whatever net effects are measured to the intervention will be certain. However, the external validity and indeed utility of the results are another matter. According to the *Evaluation Plan*:

> The impact estimates will also be externally valid, so that estimates measure without bias what the mean impact of BOND's interventions would be for the nation as a whole were they implemented nationwide. The external validity of the impact estimates is also based on the study design—the fact that the 10 BOND sites were chosen randomly allows them to represent the other areas not chosen—and the use of analysis weights whereby the BOND sites are weighted so that they represent the nation as a whole (Bell et al., 2011, p. 125).

It may be reasonable to claim that the BOND results can be generalized to an estimate of nationwide impact *at the time BOND was implemented* and under the assumption that perception of the offset among the treatment subjects is comparable to what would be achieved when the same opportunity came as part of the SSDI package as understood by all applicants/recipients and their advisors. But both the economy and the character of the SSDI and SSI caseloads are changing rapidly. As emphasized earlier, external validity is a matter of the reliability with which the results from the experiment can be used for forecasting the impact of an offset or offset-with EWIC implementation in other sites at other times. This is more what Congress may have intended in asking SSA to gather information "to adequately evaluate the appropriateness of national implementation of such a program." Nowhere in the BOND documentation is attention paid the possibility that "appropriateness" is not a matter of statistical inference about the impact of a hypothetical contemporary "scaling up" but rather whether there is adequate empirical support for key assumptions about the behavioral consequences of the kind of changes in incentive, counseling, and payments process BOND introduces.

This issue of appropriateness is touched upon in the GAO report, "Management Controls Needed to Strengthen Demonstration Projects" cited earlier (GAO, 2008). Referring to the authorizing statute for SSDI demonstration projects, section 234(b) of the Social Security Act, GAO writes:

> The authorizing statute for DI demonstration projects requires that the result derived from the projects will *obtain generally* in the operation of the disability program. While . . the BOND project . . has been designed to yield nationally representative information about the impacts of the project, the statute does not require that the results be applicable to all DI beneficiaries nationwide. However, the results should apply to a larger group of beneficiaries than just those that participated in the demonstration project (GAO 2008, 17; emphasis added).

Generously interpreted, this seems to be a call for thinking more about the external validity—and the external *utility*— of the results of the very substantial BOND investment. In any event, a first step would be to see how reliably the results from any nine sites can be used to predict the outcome in the 10th. Doing so will require more control for inter-site variation in economic conditions and other factors that mediate outcomes.

Inadequate attention to controls. The BOND plan includes effort to assess what actually happens under current WIC procedures for Stage 2 control and offset with WIC participants. But aside from the expert assessments provided by site visit teams, little effort is planned to evaluate the service environments for Stage 1 controls or to gauge the understanding of the TWP process among beneficiaries in general. While the evaluation report describes the opportunity for conversion of Ticket benefits to cash (Bell et al. 2011, pp. 30-31), nowhere is it made clear how this option will be present in EWIC or how understanding of it will be measured for the general beneficiary population.

Management feedback

The "Benefit Offset" is the headline feature of the demonstration treatment. The 50 percent benefit reduction rate, expressed as $1 for $2, is the heart of this. But it is possible that the shift to annualization of benefits will also have consequences for behavior. While the evaluation plan includes as a question "what are the lessons for national implementation of a benefit offset, future efforts to improve the design of SSDI, and broader disability policies" (Bell et al. 2011, p. 63), few clues are offered regarding the issues for which lessons might be sought. Annual reconciliation is a key feature of the income tax system, and in a sense shifting to an annualized offset underscores the idea that people are returning to the regular world of work. Suppose reconciliation were done based on information available at the time of income tax filing. How great would the difference be from what is accomplished by the August reconciliation planned under the current BOND plan?

Summary

BOND is operating on a far larger scale than the MHTS or the YTDP. However, the effort to produce results that would be in a particular sense "nationally representative" appears to have compromised other goals of the project, most notably the ability to assess the actual change in environment created by the demonstration relative to SSA operating procedure, either as currently practiced or as intended. The consequence is to muddy the connection between the focus of the demonstration and general questions of SSDI policy. Moreover, it will be difficult to link BOND results to other, related projects in the formulation of evidence-based policy.

Building links for such connections is the topic of a later section of this report. The next section considers treatment of process analysis in evaluation guides produced by other government agencies.

4. Process Analysis in Non-SSA Evaluation Guides

In this section the process analysis components of sample evaluation guidance documents from outside the SSA are briefly reviewed to discover the process analysis concepts promoted elsewhere. The review is structured around the six major process analysis points introduced in Section Two.

Three evaluation guidance documents are considered: (1) the 2012 revision of the evaluation guidelines published by the U.S. Government Accountability Office; (2) the United Kingdom Treasury's 2011 "Magenta Book"; and (3) the World Bank's 2011 impact evaluation handbook.

Many others are available[10]; however, these three are representative of the genre. The sponsoring agencies are particularly important because of the role each plays in promoting (and in some instances compelling) evaluation work by other agencies.

The Government Accountability Office

Background. The Government Accountability Office (GAO) is an agency of the United States Congress responsible for audit, evaluation, and investigation of Federal government activities. Founded in 1921 as the General Accounting Office, the name was changed in 2004 in recognition of a mission that extends beyond auditing functions and includes both program evaluation and contributing to general government efforts to improve effectiveness as reflected in the Government Performance and Results Act of 1993 (GPRA) and the GPRA "modernization" legislation of 2010. GAO's responsibilities require attention to methodology, and *Designing Evaluations*, an update of a manual printed in 1991, is intended to provide "a guide to successfully completing evaluation design tasks" (GAO 2012, p. 1).[11] Much of the material in the guide (called *Designing* here for convenience) reflects and is drawn from the work of a cross-departmental "Evaluation Dialog" organized by the Office of Management and Budget in connection with deployment of the Bush Administration's Program Assessment Rating Tool (PART) (Shipman 2006).

Summary. *Designing* begins with a definition: "A program evaluation is a systematic study using research methods to collect and analyze data to assess how well a program is working and why" (p. 3). "Program" is broadly defined as an activity, project, function, or policy with "an identifiable purpose or set of objectives" (p.3). The report is targeted both at the agency's own auditor/evaluators and at other agencies charged by Congress through GPRA and other legislation with assessing program performance and impacts. Thus, in many instances GAO evaluates the evaluations of other agencies, and *Designing* in a sense provides the template against which such efforts are audited. GAO's 2008 report on SSA demonstration projects, including BOND, is an example of this sort of work (GAO 2008).

The GAO draws an important distinction between program evaluation and its close relatives, performance measurement and monitoring. Performance measurement and monitoring involve assessment of a program's achievement relative to goals, standards, or expectations. Evaluations attempt to establish the magnitude and causes of impact, the net effect of a program compared to alternatives. An evaluation design begins with specification of a program's goals, logic and evaluation issues. Logic is the theory that connects whatever the program attempts to do with intended effects. The GAO's concept of process analysis emphasizes what programs do—"how well authorized activities are carried out and reach intended recipients." "A process evaluation," *Designing* claims, "can be an important companion to an outcome or impact evaluation by describing the program as actually experienced" (p. 15). The contribution of process analysis changes over time. Early in program implementation process provides important management

[10] See, for example, the Department of Justice practice outlined in Department of Justice Bureau of Justice Assistance (DOJBJA)(2012) and the Administration for Children and Families' excellent *Program Manager's Guide to Evaluation* (OPRE 2010). The European Commission is promoting evaluation of various European Union programs through similar guidance; see, for example, Morris, Schönhofer, and Wiseman (2012).

[11] Citations in this section follow the convention established earlier in the report: Page numbers reported alone refer to the first preceding general citation.

feedback; later "a process evaluation can be an important companion to an outcome or impact evaluation by describing the program as actually experienced" (p. 15). Virtually nothing is said of how the focus of process evaluation should be identified.

After discussing linking process to assessing the achievement of intended outcomes, *Designing* ventures toward investigating causality. The first step in GAO's procedure is to study differences in outcomes—before and after an intervention or across locations or groups— and to seek correlation with variation in external environment, implementation, or both. The next step is to consider an impact study and develop a plan for constructing a counterfactual. Both quasi-experimental and experimental (i.e., randomized control) designs are discussed, with emphasis on the difficulties attendant on each. Here the difficulty with the differences between the two audiences for the document shows up clearly—for example, "a true experiment is seldom, if ever, feasible for GAO because evaluators must have control over the process by which participants in a program are assigned to it, and this control generally rests with the agency" (p. 40). "However," the report continues, "the GAO does review experiments carried out by others."

Following review of methods for assessing causality, *Designing* returns to process briefly as a component of "comprehensive" evaluation: "Although this paper describes the process and outcome evaluations as if they were mutually exclusive, in practice an evaluation may include multiple design components to address separate questions addressing both process and outcomes" (p. 45). The guide concludes with a review of "methodological challenges" encountered in practical application of evaluation techniques, especially when programs of concern are part of complex systems in which many factors influence outcomes of interest.

Review. *Designing* is a well-written overview with admirably succinct discussion of various aspects of evaluation methodology. Evaluation methodology fills textbooks, so of course any summary will leave details out, but what is told here is told well. From the perspective I developed earlier, however, certain shortcomings are important.

First, as is common, the GAO presentation does not distinguish well between what a program is intended to deliver to its targets and the production process that leads to that end.

Second, despite the eventual discussion of causality, *Designing* is very much rooted in the literature and procedures for monitoring outcomes of the program under study. Little attention is paid the problem of comparing the effect of the program on the opportunities experienced by target groups to alternatives, including previous operating procedures. The GAO's perspective seems to be that it is internal validity that counts: "Confidence in conclusions about the program's impacts depends on ensuring that the treatment and comparison groups' experiences remain separate, intact, and distinct throughout the life of the study so that any differences in outcomes can be confidently attributed to the intervention. It is important to learn whether control group participants access comparable treatment in the community on their own" (p. 43). The implication of my analysis is that the important issue is *what* members of the control group have access to; utilizing such options, like utilizing options presented by the innovation under study, is an outcome.

The discussion of change over time in what a program accomplishes is limited in *Designing* to a distinction between implementation and "seasoned" phases of program operation.

Third, GAO's lingering audit perspective is particularly evident in failure to address external validity and the evidence base, both of which are crucial if evaluation results are to be useful in the future.

Her Majesty's Treasury

Her Majesty's Treasury (Treasury, for short) is the United Kingdom's economics and finance ministry. As such, the role played in general governance is considerably broader than that of the Treasury Department in the United States and includes many of the functions of executive agencies such as the Office of Management and Budget and congressional agencies such as GAO and the Congressional Budget Office. Like all other UK ministries, Treasury is headed by a representative of the governing party (or, as currently, parties in coalition), the Chancellor of the Exchequer. Treasury is managed by a Permanent Under-Secretary (PUS), generally referred to as Permanent Secretary. The Permanent Secretary is a non-political civil service head who holds the position for a number of years (thus "permanent"), as distinct from the changing political Chancellor to whom he or she reports and provides advice. The current PUS, Sir Nicholas Mcpherson, is an active advocate of experimental evaluation of policy alternatives.

Two documents set out Treasury's template for policy development. The first, the *Green Book* (henceforth *GB*) (HM Treasury 2006) presents the government's general framework for project, program, and policy evaluation. (Programs in the British lexicon are groups of projects (p. 1); policies presumably guide choice among and character of programs and, within programs, projects.) The framework is cast in stages, moving from rationale and objective determination through options appraisal and choice, implementation, and evaluation. This policy cycle is known by an acronym formed from the boxed stations: ROAMEF. It is reflected in Figure 5. *GB* discusses requirements at each station of the cycle; government departments seeking Treasury funding attempt to cover all these bases.

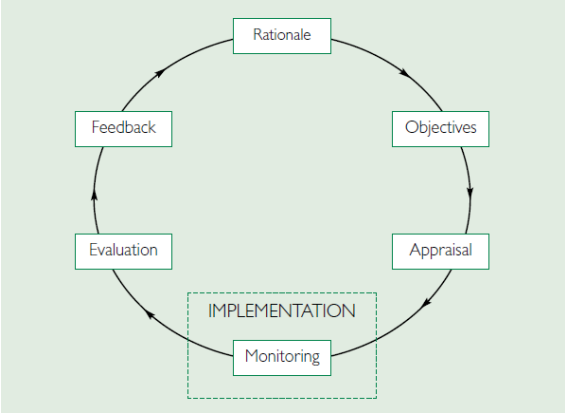

Figure 5: The ROAMF Policy Cycle

The second, the *Magenta Book* (henceforth *MB*)(HM Treasury 2011) is the guidance specific to best evaluation practice for departments. The *MB* clarifies that the stepwise depiction of the ROAMEF cycle implies a sequence that rarely holds. In practice,

> . . the process is often iterative and there are significant interdependencies between the various elements. . . . In addition, evaluations can play a role in the policy development process – through, for instance, the use of pilots and trials – implying the presence of (potentially numerous) feedback loops at different stages of the cycle. Therefore whereas the simple ROAMEF policy cycle shows that an

> evaluation will take place after the policy has been implemented, evaluations can, in fact, occur at practically any other time. And importantly, decisions affecting and relating to any evaluation will almost always be taken much earlier in the policy process. . . .[W]hat might seem minor aspects of the way a policy is formulated or implemented can have significant impacts upon the ability to evaluate it rigorously. It is important, therefore, to ensure that evaluation is considered and planned at the same time as the policy is being formulated so that these links can be recognised and accounted for. (p. 15)

Treasury's emphasis on planning evaluation in conjunction with policy planning is missing from the GAO guidance, although "pilots and trials" do come under scrutiny in some GAO assessments. But the *MB* perspective on process misses the counterfactual measurement issue. According to the *MB*, process evaluation asks how a policy was delivered, not what was its net impact on experience. "In general, process-related questions are intentionally descriptive" (p. 18). This contrasts, according to *MB*, with impact evaluation, which asks "What difference did the policy make?" (p. 17). The guidance proceeds with a very comprehensive review of procedures for building impact evaluation into policy design, field study of policy implementation, and impact evaluation execution, including quasi-experimental and experimental designs.

The key *MB* problem lies in treating process and impact evaluations as if the concept of counterfactual arises only in the context of impact assessment and then principally in relation to outcomes, not inputs. This perspective is reflected at the end of chapter 2 ("Identifying the right evaluation for the policy"). There the *MB* states:

> There is then the additional consideration of what sort of answers process and impact evaluations can provide. This chapter has portrayed the answers from process evaluations as more descriptive, and the answers from impact evaluations as more definite and in some sense "robust". This is because good impact evaluations attempt to control for all the other factors which could generate an observed outcome (that is, they attempt to estimate the counterfactual). But again, the distinction between the two is not as simple as this suggests. (p. 21)

What the *MB* fails to recognize is that assessing the counterfactual *process* is essential to full understanding of the sources of whatever outcome impacts are observed.

The World Bank

Interest in sophisticated assessment of policy in developing countries has grown significantly in recent years, propelled by philanthropic organizations, international aid agencies, and recently by aggressive and well-funded advocacy by the Abdul Latif Jameel Poverty Action Lab (J-PAL) at the Massachusetts Institute of Technology and the J-PAL network of affiliates (Banerjit and Duflo 2011). As a premier funder of both human and physical development investments, the World Bank has a long-standing interest in both the methodology and practice of evaluation. For over a decade several groups within the Bank have organized workshops for civil servants and planners on evaluation techniques under the general title of "Turning Promises to Evidence." *Impact Evaluation in Practice* (Gertler et al. 2011)(henceforth *IEP*) pulls the materials for these

workshops into a textbook. Given the importance of the Bank's activities and the high quality of Bank staff, *IEP* is worthy of study as an example of attention to process analysis.

IEP is the longest of the three guidance documents reviewed here. Like *Designing* and *MB*, *IEP* is in part intended to make the case for evaluation, but as the title indicates the focus is almost exclusively on impact assessment, which the authors see as part of a broader agenda of evidence-based policy making. The importance of prospective evaluations, engineered as part of program implementation, is a major theme, as was also the case in the *MB*. Much evaluation guidance begins with discussion of what happens in the intervention and then moves from process assessment to impact evaluation. *IEP* starts by explaining why estimation of the unobserved counterfactual, i.e., what would have happened to program participants in the absence of the program, is the central impact evaluation issue. Other guidance documents approach classical evaluation gingerly, after going through alternatives. *IEP* goes straight to the importance of estimating the counterfactual in impact analysis. RTC is justified as producing in most instances "an excellent estimate" of the counterfactual. Other methods are interpreted as attempts to replicate RCT estimates.

IEP does consider the nature of the intervention, referring, as do both *Designing* and *MB*, to the logic model or theory of change underlying the intervention. The Bank's conceptual scheme for innovations, called the "Results Chain," is quite similar to the demonstration model developed in my Figure 3. In the *IEP* results chain (see Figure 6), what I term the "interface" in my model occurs at the transition from outputs to outcomes. *IEP* is exceptional among evaluation guidance documents in explicitly recognizing the interaction between demand and supply—the response of the targeted population—as an intermediate step in the move from inputs to the outcomes of ultimate ("final") concern.

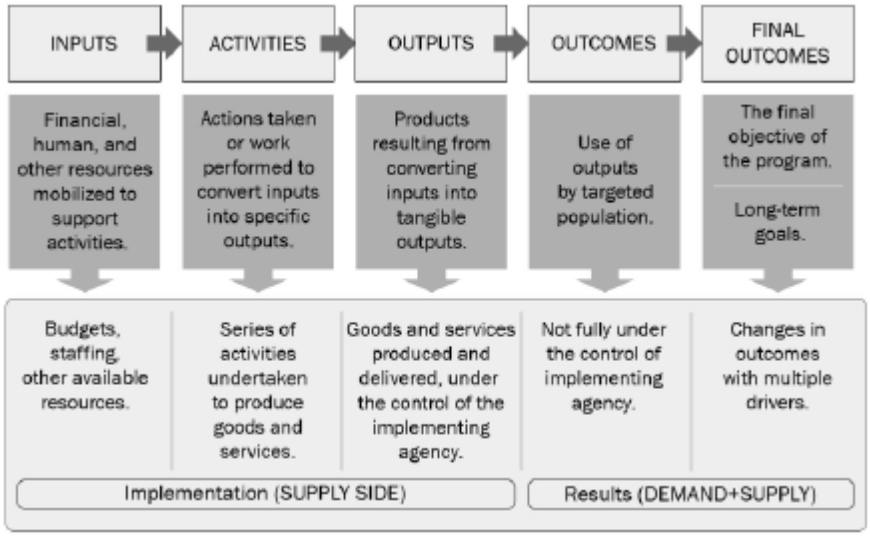

Figure 6: "Results Chain" from *Impact Evaluation in Practice* (Gertler et al. 2011, p. 25)

There is much to study in the Results Chain. *IEP* recognizes that attempting to measure and monitor every aspect of the logic model would be a waste of resources even if feasible. The challenge is to find "SMART" indicators: Good ones are Specific (as close to the information required as possible), Measurable (readily obtained), Attributable (linked to the project's efforts), Realistic (capably of being obtained in timely fashion, with reasonable frequency, and at reasonable cost), and Targeted to the objective population (p. 27).

Consistent with its emphasis of the problem of counterfactual estimation, *IEP* defines the internal validity of an impact evaluation as occurring if a "valid comparison group" is used, i.e., the

comparison group represents the "true counterfactual" (p. 54). External validity is defined very narrowly: "An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units. The results are then generalizable to the population of eligible units" (p. 54). The authors seem to mean by this the contemporary population from which the experimental sample is drawn. But external validity encompasses a more general capacity for use of results for forecasting. The Bank's notion of external validity cannot be tested, but ability to forecast can. I discuss this issue earlier in connection both with BOND and MHTS.

Figure 6 involves process as well as outcome. Despite its explicit attention to counterfactual estimation, *IEP* in fact pays little attention to process analysis, adopting a perspective similar to that of *MB*. Process evaluation is, in this view, one thing; impact is something else:

> [P]rocess evaluations focus on how a program is implemented and operates, assessing whether it conforms to its original design and documenting its development and operation. Process evaluations can usually be carried out relatively quickly and at a reasonable cost. In pilots and in the initial stages of a program, they can be a valuable source of information on how to improve program implementation. (p. 17)

IEP does acknowledge that process evaluations are of some use; along with qualitative and monitoring data,

> . . . process evaluations are needed to track program implementation and to examine questions of process that are critical to informing and interpreting the results from impact evaluations. In this sense, impact evaluations and other forms of evaluation are complements for one another rather than substitutes. (p. 15)

But what should be in the process evaluation to make it complementary? What features would be, in the *IEP* scheme, "SMART"? The problem seems to be the presumption that the nature of the innovations that concern the Bank is such that members of the control group have no opportunities comparable to those generated by the intervention to be evaluated or, if they do, the opportunities are representative of those available to the eligible universe. No thought is given the reliability of this presumption.

Summary

All three of the guidance documents discussed in this section offer valuable insights into evaluation procedures. All three have shortcomings with respect the assessing the impact of an innovation on the environment of the target population—what is, from my perspective, the essential objective of process analysis. The consequence is that evaluations following such guidance may be deficient for assessing replicability. They also may fail to produce the information on process impact that is important for building the relevant knowledge base for evidence-based policy.

I explore the link to evidence-based policy further after reviewing three examples of process analysis.

5. Process in Action

Examinations of process that emphasize assessing the net effect of interventions are increasingly common in the literature. This section illustrates such efforts with three examples. The first, a study of prison-based correctional programs for inmates involved with drugs, is an example of careful study of site-to-site variation in treatment fidelity. The second study uses welfare-to-work experiments as a source of information on variation in the control environment. The third, a study of the impact of variation in treatment for welfare-to-work programs on outcomes, illustrates the utility of combining multi-site demonstrations and multiple demonstrations to identify the impact of variations in the character of the demonstration-created target group environment.

Process Analysis in a Family of Correctional Programs

I emphasize throughout this report the importance of finding measures of the change in environment of target group members that is created by demonstrations. In "Improving process evaluations of correctional programs by using a comprehensive evaluation methodology" Jeffrey Bouffard, Faye Taxman, and Rebecca Silverman (2003; henceforth BTS) report results of multiple methods of fidelity assessment for seven jail-based "therapeutic communities" (TCs) for "drug-involved" offenders. While there are many variants, the normal TC model calls for establishing a community of participants in a separate living environment. The community is then guided in developing a positive peer culture to support acquisition or re-acquisition of values and behaviors supporting reduction of drug abuse. Randomized controlled trials indicate that TCs can, when combined with community-based care when incarceration ceases, significantly reduce both drug use and recidivism rates.

The details of the TC model are well established (De Leon 1994). BTS are concerned with assessment of the fidelity of actual TC implementation to the model. They examine program operations in seven sites. Three approaches to assessing target group experience are considered. The first is staff interviews involving specific questions concerning client selection, treatment modalities, drug testing practices, the application of sanctions for rules violations, and the presence or absence of community treatment for those leaving incarceration. The second is review of administrative reports concerning various program aspects such as frequency and results of drug testing, rules infractions and the consequence of sanctions, and the movement of target group members from residential to community treatment.

The third assessment is termed "systematic social observation," or SSO. "SSO attempts to record, in an objective, quantifiable manner, the characteristics of a given social environment" (p. 151). Basically SSO is a data collection protocol for on-site investigators based on features deemed important by the TC logic model. SSO procedures are systematic and replicable. The SSO instrument for the study reviewed here had five components, covering program emphasis, topics covered in materials presented to the target group, types of media used, treatment "style," and the ways in which the TC peer groups function. Some 66 separate measures were used.

The authors discover many things. The first is that despite being nominally the same, the actual TC programs differed substantially in emphasis across sites. Second, in general the programs failed to implement core components of the TC model, in particular encouraging target group

members in "contemplation of change" and "self-work" (p. 156). Third, the SOS data generally provided a far more complete picture of TC implementation and fidelity to the TC model than did either of the other techniques; indeed, in several instances information provided by stakeholders was proved wrong and the interpretation of information from management reports was demonstrated to be at least questionable. The implication is that one-time site visits with focus on stakeholder interviews and review of a few management indicators may be questionable as a source of information on treatment accomplishment. The big question, unanswered by BTS, is what the impact of this variation might have been on the outcomes—recidivism and drug dependence—of primary concern.

As described by BTS, SOS is very labor intensive and probably too expensive for replication in the context of SSA-scale experiments. But the elements of SOS are worth consideration as an example of translating logic models into intended experience, and then looking for measures of what is achieved.

Looking More Closely at Controls

For a variety of reasons, there are more randomized trials of work-oriented welfare demonstrations than for any other domain of social policy. Indeed, the landmark social policy RCT is the famous New Jersey negative income tax experiment of the late 1960s, which focused on what would happen if "welfare" was extended to two-parent poor families. Over the subsequent half-century, dozens of such experiments have been undertaken, with generally increasing sophistication. Following common practice, the term "welfare" refers to a means-tested benefit program, most frequently the Temporary Assistance for Families (TANF) program or TANF's predecessor, Aid to Families with Dependent Children (AFDC). A welfare-to-work program is some sort of intervention intended to raise employment rates among welfare recipients. An obvious next step, frequently taken, is to attempt some sort of synthesis of the results. I turn next to two of these syntheses. Both offer insights relevant to my major themes.

The first is a paper by David H. Greenberg and Philip K. Robins (2011; henceforth GR) entitled "Have Welfare-to-Work Programs Improved Over Time in Putting Welfare Recipients to Work?". GR assemble results of 21 random assignment evaluations conducted in various locations about the U.S. between 1983 and 1998. All included provision of some "active intervention" such as job search requirements, public service employment, remedial education, or job training. All were in some way "mandatory" in that failure to participate could bring loss of benefits. Employment in the seventh quarter after random assignment was taken as the major outcome of concern. The authors increase the number of observations by counting each site in multiple-site demonstrations as a separate program and also by treating observations on outcomes for adults in different family circumstances (one- versus two-parent) as if for separate programs. Counted this way, the total number of programs observed can reach as high as 73; actual sample size varies because a full information set is not available for each trial.

The weighted average impact of the interventions on seventh-quarter employment rates across the entire time span of the studied demonstration is 2.6 percentage points, from 37 percent for controls to nearly 40 percent for members of the treatment groups. What interests GR is that the estimated impact of treatment is effectively uncorrelated with the time in which the demonstrations occurred. This lack of trend is disturbing, for it could be taken to suggest that

policymakers have not gotten better at promoting employment. It persists even when estimated impacts control for area unemployment rates, target group characteristics, and other (external) environmental characteristics.

GR propose an alternative to this lack-of-learning hypothesis: They argue that over time services available to controls have increased, so that the net effect of demonstrations on opportunities has diminished. The largest change appears to be in reported incidence of job search activities among controls. If the GR regressions are to believed, had the difference between treatment and controls in job search activities stayed constant at levels reported in early demonstrations, the average demonstration impact on employment would have doubled. They fault the failure of implementing welfare agencies to test more significant departures from the *changing* scene of what is already available. To this end, they suggest that, "in future experiments, evaluators need to play a more significant role in the design of the treatment." Such participation would, they claim, be more likely to produce demonstrations with "sizable employment impacts" with potential to be "cost effective" (p. 919).

The GR analysis is important because of its novel emphasis on what has occurred with controls—an emphasis consistent with my process analysis framework. However, there are many problems. Treatment of sites within a common demonstration framework as independent separate observations (such as differences between one- and two-parent cases) ignores the likely presence of unobserved commonalities within each demonstration "cluster." The GR paper treats outcomes—participation in particular program components—as inputs. For example, the incidence of sanctions is taken as a reflection of the rigor of application of work requirements. But, as is pointed out in the BTS report noted earlier, low levels of sanctions in mandatory programs may reflect target group response to clear communication of requirements and consistency of penalties. Most seriously, GR fail to consider whether measurement of control activity (or more appropriately, control opportunity) has improved over time. In the early days of RCT evaluation of welfare-to-work experiments, little attention was paid to characterization and measurement of input, especially on the control side. The presumption was that the treatment was different, and the focus was on outcome impact. The quality of measurement of control activity, let alone opportunity was low. Over time, the ambition of demonstrations grew, and deficiencies in control assessment may have declined.

Process in Multi-Level Impact Analysis

Arguably the best example of integration of process and impact analysis is provided by Howard Bloom, Carolyn Hill, and James Riccio in their 2003 paper "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments" (Bloom et al. 2003; henceforth BHR). Like GR, BHR take advantage of the accumulation of welfare-to-work program evidence during the last two decades of the 20th century to study the impact on outcomes of variation in program management, target group characteristics, target group opportunities, and the economic environment in which programs are implemented.

In the tradition of meta-analysis, GR attempt to include all welfare-to-work demonstrations that occurred within the time period they consider, excluding some only because of important missing information. In contrast, BHR only use data from three evaluations done by the same firm (MDRC, for which the authors worked as employee or consultant) that were conducted in

the context of federal welfare regulations under rules established by the Family Support Act of 1988. Like GR, BHR treat local offices as separate sites, thus producing impact over time in 59 locations. But unlike GR's, the BHR analysis exclusively involves impacts for single parents. While the dataset used by BHR is more restricted, the great advantage they have is availability of substantially better information on the environment to which treatment and control groups were exposed.

The central outcome for the analysis is employment, measured by BHR as sample members' total earnings over the two years following random assignment. A two-level hierarchical model of earnings determination is estimated. In the first level, outcomes are estimated for each office, with target group characteristics and control/treatment status as right-hand variables. This provides an estimate of treatment impact at each office conditional on local target group characteristics. In the second level, each estimated office impact is regressed on measures of program implementation, activities available to the target group, and the external economic environment. Combining the two levels allows estimates of the impact of each factor on outcomes.

Of particular interest here is the way BHR represent and measure treatment and control environments. Program practices are represented by results of staff and supervisor surveys and include things like "emphasis on quick job entry for clients," "emphasis on personalized client attention," "closeness of client monitoring," and caseload size. In addition to direct measurement of individual assessment of these features, the model includes measures of within-site variation in assessment and differences in assessment between supervisors and line staff; these measures are included to give a sense of presence or absence of a common office vision of program objectives.

Much like GR, BHR consider participation in common activities as indicators of treatment beyond office atmosphere. Differences in treatment and group participation rates are estimated for each office for job-search assistance, basic education, and vocational training. It is important to note that actual participation in any of these activities is not part of the level one regression. Rather, the difference in prevalence of such activities for treatment and control group members at each site is treated as an indicator of emphasis. The model I use in this report would call these prevalence rates instruments for availability.

The results of the BHR analysis provide important evidence of the effect on outcomes of variation in interface as well as economic environment. Emphasis placed on quick job entry and personalized service significantly affects treatment impact on earnings; caseworker caseload size is inversely related to impact. Offices with high participation in basic education produce lower earnings impact, and the average site unemployment rate over the two years for earnings measurement is inversely associated with net effects.

There are shortcomings in the BHR analysis. Most treatment characteristics are measured only at one point in time, through surveys. It is not clear from the published report how many of the scales of program implementation were measured separately for treatment and control observations; it seems unlikely that the strong employment orientation of some offices did not "bleed over" to affect behavior of controls. Nevertheless, the multi-stage framework is a model for other analyses and the outcome provides strong evidence of the importance of process

assessment. The BHR conclusions, while based on programs different from those fielded by SSA, appear to have more general validity. They write:

> [W]e emphasize that our research was possible only because of the careful, comprehensive, and consistent data collection efforts of the experiments that we pooled and the broad range of circumstances that they represent. Thus, as social scientists and policy researchers develop their research agendas and as government agencies and foundations make their future research funding plans, we urge them to emphasize a long-run strategy for accumulating program knowledge based on:
>
> 1) random assignment experiments that make it possible to obtain valid and reliable estimates of program effectiveness,
>
> 2) multi-site experiments that reflect the existing range of natural variation in program effectiveness,
>
> 3) careful specification of the likely determinants of program effectiveness based on social science theory, past empirical research, and experiential knowledge of practitioners,
>
> 4) equally careful and consistent measurement of these hypothesized determinants across studies and sites, and
>
> 5) adequate support for and attention to quantitative syntheses of this information. In this way we believe that the most progress possible can be made toward unpacking the "black box" of social programs and thereby acquiring the information needed to improve them. (p. 572)

6. Process Analysis and Evidence-Based Policy

"Evidence-based policy" is widely endorsed. Determining just what evidence counts has been the focus of considerable effort, notably by the Coalition for Evidence-Based Policy (CEBP) and related organizations. This section argues that the attention paid by CEBP to the internal validity of demonstration evaluations needs to be complemented with more attention to what process analysis contributes to external validity. My concern is linked to other recent work urging more attention to context and treatment/control differential in evaluating "scaling up" demonstrations.

Obama Administration Initiatives

Attention to the meaning and promotion of "evidence-based policy" has grown in recent years, in significant part because of initiatives of the Obama administration (Haskins and Baron 2011). By the end of 2010, the administration had six new evidence-oriented initiatives under way. These included Health and Human Services (HHS) initiatives for teen pregnancy prevention and improved family functioning, a collection of Department of Education "Investing in Education" (i3) projects, two joint Education-Labor initiatives related to workforce development, and creation of a Social Innovation Fund to provide grants via intermediary organizations to local organizations engaged in the conduct and evaluation of evidence-based programs directed to certain broad areas of social policy. Proposed budgets for FY 2011 and FY 2012 included

requests for various additional evaluations "that have the potential for strong study designs and that address important actionable questions or strengthen agency capacity to support such strong evaluations." (OMB 2011, 83)

The *Analytical Perspectives* section of the FY 2014 budget continues and expands this emphasis on evidence:

> In the area of evaluation, the Administration has moved to adopt a multi-tiered approach to evidence-based funding for new grant-based initiatives targeted towards education interventions, teenage pregnancy prevention, social innovations, home visitations for new parents, workforce interventions, and science, technology, engineering, and math programs. The initiatives offer the most funding to programs and practices supported by the strongest evidence. Programs with some, but not as much, supportive evidence also receive significant funding, the condition [sic] that the programs will be rigorously evaluated going forward. Over time, the Administration anticipates that some second-tier programs will move to the first tier as they prove more promising and cost-effective than other programs. Finally, agencies are encouraged to innovate and test ideas with strong potential—ideas supported by preliminary research findings or reasonable hypotheses. At all levels, it is important to build implementation evidence into this multi-tiered approach so that we understand how best to scale successful programs and to create more and better program options. (OMB 2012a, 92)

This statement has many important features. One is the reference to a "multi-tiered approach," with initiatives supported by the "strongest evidence" designated as "first tier." A second is the reference to moving innovations from second to first tier as evidence of effect and efficiency accrue. Yet another is the importance attached to building "implementation evidence" into evaluations as an aid to scaling successful programs—i.e., ensuring results are externally valid and therefore can be replicated with comparable impact. The commitment was reinforced by a subsequent memorandum from the Office of Management and Budget (OMB) calling on federal agencies to "demonstrate the use of evidence throughout their Fiscal Year (FY) 2014 budget submissions" (OMB 2012b).

The Coalition for Evidence-Based Policy

The content of White House/OMB strategy for evidence-based policy has been significantly influenced by the work of an interest group outside government, the Coalition for Evidence-Based Policy, or CEBP (the CEBP is cited specifically in the OMB memorandum).[12] According to its website, this organization "seeks to increase government effectiveness through the use of rigorous evidence about what works." Part of that influence is generated by an effort to clarify what is meant by "top tier" social program models (policy "interventions") and to identify interventions that meet, or come close to that standard. The standards are applied by panels of experts to evaluation results.

[12] See www.coalition4evidence.org. Unless otherwise indicated, information on the activities of the Coalition for Evidence-Based Policy presented here is taken from this website and the "sister" website www.evidencebasedprograms.org (accessed 17 June 2012).

The CEBP advocacy strategy includes promotion of specific standards for assessing intervention evaluations and for identifying the "top tier" interventions the organization sees as worthy of widespread adoption. The beginning point is the importance attached to classical evaluation:

> The Coalition advocates many types of research to identify the most promising social interventions. However, a central theme of our advocacy . . . is that evidence of effectiveness generally cannot be considered definitive without ultimate confirmation in well-conducted randomized controlled trials.

The Coalition posts a checklist for reviewing RCTs to assess whether the evidence produced is "valid." (CEBP 2012) The checklist is exclusively directed toward ensuring the evaluation has internal validity, although an appendix addresses the issue of the number of RCTs "needed to produce strong evidence of effectiveness." (CEBP 2012, 6) "To have strong confidence that an intervention would be effective if faithfully replicated," the CEBP checklist states, "one generally would look for evidence including":

- ✓ "The intervention has been demonstrated effective, through well conducted randomized controlled trials, in more than one site of implementation . . .
- ✓ "The trial(s) evaluated the intervention in the real-world community settings and conditions where it would normally be implemented . . .
- ✓ "There is no strong countervailing evidence, such as well-conducted randomized controlled trials of the intervention showing an absence of effects." (CEBP 2012, 6)

Thus top-tier interventions satisfy the validity checklist and are supported by "at least two well-conducted trials or, alternatively, one large multi-site trial." Near-top tier interventions satisfy the checklist, but lack replication. Beyond the CEBP framework, some descriptions add a third, lower tier for interventions that show promise on the basis of some research findings but lack rigorous confirmatory evidence. The federal agency responsible for operating the Social Innovation Fund (SIF), the Corporation for National and Community Service (CNCS) thus differentiates among interventions supported by "preliminary," "moderate," or "strong" evidence. Strong "means evidence from previous studies on the program, the designs of which can support causal conclusions (i.e., studies with high internal validity), and that, in total, include enough of the range of participants and settings to support scaling up to the state, regional, or national level (i.e., studies with high external validity)" (CNCS 2012, 10).

With respect to the three SSA demonstrations I focus on in this report, the CEBP and CNCS standards indicate that BOND is designed to produce top-tier evidence, and that the MHTS fits the standard as well. The YTDP demonstrations are sufficiently diverse so that only second-tier evidence is likely to be produced, but replication of the successful YTDP models would have first-tier implications.

The efforts by the CEBP and the CNCS are commendable. However, from the perspective of my discussion here the absence of reference to process in the standards is puzzling. As indicated by the checklist presented above, the standards refer almost exclusively to the internal validity of a program evaluation; external validity is viewed as accomplished by replication of the treatment, preferably "in the real-world community settings and conditions where it would normally be

implemented." There is no component of the checklist that relates to the description of process achievement or the experience of controls against which the treatment is measured.

The W. T. Grant Initiative

The failure of the CEBP/SIF standards to include adequate consideration of process data has generated some response. The W.T. Grant Foundation has been particularly interested in improving understanding of the factors that influence the impact of innovations, and the implications of these moderating influences for predicting the consequences of "evidence-based" policy in education.

Much of this work has been promoted by Robert C. Granger, the Grant Foundation's president (cf. Granger 2011). The foundation focuses on interventions for children, such as the ED "i3" initiative cited above. Granger's argument is that many interventions in the education sphere have been subjected to multiple impact interventions. The problem is that impact estimates vary, and often a meta-analytic conclusion of "positive impact" is the product of a combination of a few studies detecting major net effects and an embarrassing number of trials with small or no consequence. Scaling up requires knowing not just that the intervention can have an effect in some circumstances, but finding out *just what those circumstances are*. All too often the detail—the process analysis—is missing. This shortcoming makes it difficult to sort out reasons for the variance in outcomes.

Granger has suggested that the next round of i3 projects pay more attention to process—drawing a clear contrast within each initiative between treatment and control groups (Granger 2011b). Accumulating more experimental evidence, if combined with better information on treatment, control, and context, could, he argues, "help us move beyond 'what works' to learn why and under what conditions programs are effective" (Granger 2011b, 3) More detail is needed, of course, on the type of information needed to achieve this end. My final section now turns to how the CEBP checklist can be enhanced to address Granger's concerns in the context of Social Security-oriented intervention evaluations.

7. Conclusion

The introduction to this report poses the question, "What are the essential elements of process analysis?" Logically, the conclusion is a list, applicable *ex ante* by the SSA staff members charged with evaluating "basic demonstration project development and design issues" (SSA 2010, 19). The same list should be relevant to assessment *ex post* of demonstration results. The conclusion of this survey is that the six point framework I developed in Section 2 yields a checklist of 10 elements to consider: (1) target, (2) treatment, (3) circumstance, (4) perception, (5) measurement, (6) production, (7) trajectory, (8) location, (9) choices, and (10) connection.

(1) Target. What are the outcomes that the demonstration is intended to affect? What is the target group for these consequences?

(2) Treatment. What is the demonstration intended to deliver for the target group? What is the theory, or logic model, behind the hypothesized connection between treatment and outcome?

(3) Circumstance. What is the baseline, or control environment against which this treatment is to be compared? What is supposed to change about the environment of the target group as a consequence of the demonstration? What alterations are important to the logic model for the demonstration, and which ones matter most?

(4) Perception. What are baseline treatment group perceptions of their environment? How are these expected to differ with the treatment? What are the likely consequences of the experimental nature of the demonstration for perception, as compared to what might be expected with routinized implementation? Are these consequences a threat to the external validity of the demonstration results?

(5) Measurement. What instruments are available for assessing the environment features identified by the logic model as important and the way in which the environment is perceived both for treatment and control groups?

(6) Production. What is the management model for producing the treatment interface? Is there feedback to management from measures of program fidelity? Is it possible to use measures of interface achievement for both management and impact evaluation? Will operations in place produce the desired control, or counterfactual environment?

(7) Trajectory. How is measurement conducted over time? How much will be recorded about changes in treatment and control environment through the life of the experiment?

(8) Location. If the demonstration is conducted in multiple sites, why? How does this justification affect the process analysis plan? What intersite differences may be anticipated in both treatment and control circumstances?

(9) Choices. Demonstration process has many dimensions. Given that resources are scarce, every process analysis plan incorporates choices about where to focus measurement effort and how much effort to exert. A plan should identify and justify the choices made in light of consequences for both the internal and external validity of the project's impact assessment. Here, as in identifying critical features of the demonstration's interface, it is important to prioritize. Where is the marginal opportunity actually incorporated, and where is the marginal opportunity forgone? Is the choice made efficient—i.e., producing more gain than loss—in light of the overall goals of the demonstration to provide "evidence of the feasibility and effectiveness of a new approach or practice"?

(10) Connections. The final check concerns the relationship between the demonstration under consideration and both what has come before and what might be expected afterward. This is the contribution to meta-analysis that will add to the evidence base for reform. Are any aspects of the demonstration and control processes sufficiently well-defined so that when impact is established, the results can be credibly combined with estimates from earlier work? What about replication in other locations? If the demonstration plan fails to identify the impact of the innovation on process, the external validity and, indeed, the external *utility* of the results will be at best questionable. Thus, much is at stake in process analysis development.

8. The Ten-Fold Path

My concluding 10-item checklist can now be restated as guidelines for the process analysis component of SSA demonstrations. I limit these guidelines to demonstrations intended to assess impact of innovations on the clients or potential clients of SSA programs, not efforts intended to improve the efficiency of delivery of programs as currently designed. The three SSA interventions discussed in this report—MHTS, YTD, and BOND—all have this character. Cast as 10 "steps," the guidelines are intended for use in two ways. The first is as a checklist for demonstration planners, a response to the question posed in the introduction to my report. The second is as checklist for demonstration evaluators: What should be there, and is it?

Step One: Review the logic

Planning process analysis begins with the demonstration scheme: The target group of people, the primary outcome of interest for this target group, the treatment, and the theory of (logic model for) the innovation. The logic model specifies the change to which the target group is to be treated and the outcome anticipated as a result. Assessing the change in outcome is the province of the impact analysis. Assessing the change in the environment the treatment achieves is the object of process analysis. Without theory, one doesn't know where to look.

Step Two: Distinguish processes

Step two is to separate the activities associated with establishing the counterfactual, i.e., developing a prediction of what would have happened to treatment group members in the absence of the innovation, from the activities associated with delivering the treatment. The means by which the counterfactual is developed are relevant to the process analysis for treatment impact insofar as the experimental environment, and not the only the treatment, influences the anticipated outcome.

Step Three: Identify the changes that count

Step three is to specify what change in the environment of the target group the demonstration treatment is expected to accomplish. This change is best expressed in terms of opportunities the demonstration affords the treatment group. Because Step three is cast in terms of change, it requires a statement of what is presumed to be true in the counterfactual environment. Change has many dimensions, and most may be irrelevant to the internal or external validity of the evaluation results. The logic model for the intervention should point to what features count in predicting the primary outcome. In demonstrations to be carried out in multiple sites, it is usually appropriate to cast Step three in terms of a single representative location. The complications and advantages of multiple locations can then be addressed from this base.

Step four: Find the measure

Step four is to develop a measurement plan for the environment the demonstration produces and the corresponding elements of the experience of the controls. This is tricky, because, as I discuss at various points, what is typically observed is not actual change in target group environment but rather mediating steps in interface delivery or mediating target group responses. The art here lies

in making the case that observed mediating activity or target group response is related to the opportunity created. For example, if, as in BOND, part of the demonstration treatment is an increase in available employment counseling services, then a mediating production response should be an increase in the number of available services relative to the number of participants. A mediating target group response that reflects the change in environment brought about by BOND might be the frequency of participant counselor meetings.

Step five: Anchor fidelity

Step four identifies the focus of process analysis. In Step five the fidelity counterfactual is defined. How will observers recognize that the treatment is being delivered as the demonstration intends? How are deviations likely to occur, and will this show up in the process measures identified in Step four? Step five is an important point of dialog with the managers responsible for treatment delivery. What will they be watching, and what is the connection between management objectives and the interface the demonstration intends to create? Can management indicators serve as process measures or at least as instruments for the interface that cannot be directly observed?

Step six: Consider connections

The sixth process analysis step is more of an ancillary activity than a discrete step *per se*. This step is to think "before, after, and alongside" the demonstration at hand and to consider its contribution to the policy evidence base. The "before" component is a look backward at antecedent studies similar in character. What is known about the process component of such work? Is the treatment achieved in such studies comparable in character and measurement to what is contemplated by the present demonstration? The "after" component concerns the likely feasibility of confirming impact estimates derived from the present demonstration by implementation of initiatives similar in achievement in the future—in other words, producing through replication "top tier" evidence. The "after" component also looks to the requirements for scaling up. Suppose the treatment is determined to be efficient at the scale of the demonstration. What are potential process impediments to general implementation, and what evidence will the process study provide concerning their importance?

SSA demonstrations are commonly conducted at multiple sites. The simplest motivation for site multiplication is sample size—no single location provides enough members of the target group to provide adequate statistical power for impact assessment. Generally, however, sites are multiplied for other purposes such as testing the sensitivity of results to variation in administrative, economic, or social environment or to enhance external validity. For process analysis, the important questions concern dimensions of variation in treatment or control environments that are relevant to the impact of the demonstration on the primary outcomes of concern. The "alongside" component calls for identifying the intersite differences that are likely to lead to variation in impact. Will the planned interface measures capture such differences?

Step seven: Think about trajectory

For a variety of reasons, the impact of the demonstration on target group experience is likely to vary over time. This can occur, for example, because those responsible for managing the

treatment gain experience and improve fidelity or because developments outside the control of demonstration management affect what happens to control group members and, in consequence, the difference between control and treatment experience. Step seven is to assess how the analysis plan identifies change in treatment and control experience over the life of the demonstration.

Step eight: Specify the minimum

This could be called the "get serious" step. Plans for process analysis can be excessively ambitious and, in consequence, raise the chance of failure to gain reliable information of any sort. Step eight is to review steps one through seven with an eye toward identifying minimally acceptable accomplishments. What, for example, might be the single best indicator of what the demonstration changes for treatment group members compared to the control or counterfactual experience? "Best" here must be defined with reference to the primary object of the demonstration.

Step nine: Consider tradeoffs

For each of the first seven steps it will be possible to identify possible accomplishments that exceed the minimum. Indeed, identifying the minimum (Step eight) probably involved culling from a larger collection of possible achievements. Step nine is to look back at what was discarded in getting to the minimum and to identify "would-if-we-could" (WIWC) items. Nothing is done without costs. The WIWC items should be ranked in terms of rough benefit/cost assessment. This establishes an expansion path for the demonstration process study. Assuming the path is followed in the design for the demonstration, though should be the best route for contraction as well, should developments require it.

Step ten: Check your hats

An evaluation strategy reflects a combination of interests, and many of the choices made in carrying out Step one thorough Step nine deal with matters of concern to other facets of the demonstration plan. Thus negotiation will be involved and the process strategy eventually adopted will reflect effort to satisfy multiple stakeholders. Think of them as represented by four "hats." One hat is worn by the treatment manager, who is concerned about how the treatment is delivered and how fidelity will be assessed. Another hat is worn by the person responsible for outcome impact assessment. In the final demonstration assessment, any conclusions regarding impact or lack of impact on the major outcomes will likely refer to process. What is the perspective of the impact evaluator? Hat three belongs to the benefit/cost accountant, who needs numbers on production process for both treatment and controls. Hat four, a big one, belongs to the sponsoring agency. From this final perspective what is critical about all choices is the consequence for achieving demonstration goals.

References

Banerjee, Abhijit, and Esther Duflo. 2011 *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Bell, Stephen, Daniel Gubits, David Stapleton, David Wittenburg, Michelle Derr, Arkadipts Ghosh, Sara Ansell, David Greenberg. 2011. *Evaluation Analysis Plan*. Baltimore, Maryland: U.S. Social Security Administration.

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.

Bouffard, Jeffrey A., Faye S. Taxman, and Rebecca Silverman. 2003. "Improving process evaluations of correctional programs by using a comprehensive evaluation methodology." *Evaluation and Program Planning*, 26, 149-161.

Coalition for Evidence-Based Policy. 2012. *Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence*. Washington: The Coalition. URL: http://toptierevidence.org/wordpress/wp-content/uploads/Top-Tier-Checklist-for-Reviewing-RCTs-Updated-Jan10.pdf. Accessed 13 June 2012.

Corporation for National and Community Service. 2012. *Overview of Funding Opportunity: Social Innovation Fund*. Washington: The Agency. URL: www.nationalservice.gov/pdf/12_0210_sif_nofa.pdf. Accessed 17 June 2012.

De Leon, George. 2004. "The Therapeutic Community: Toward a General Theory and Model" in Tims, Frank M., George De Leon, and Nancy Jainhill, editors, *Therapeutic Community: Advances in Research and Application.* Rockville, Maryland: National Institute on Drug Abuse, 16-53.

Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation (OPRE). 2010. *The Program Manager's Guide to Evaluation*, 2nd ed. Washington: The Agency. URL: http://www.acf.hhs.gov/programs/opre/other_resrch/pm_guide_eval/reports/pmguide/program_managers_guide_to_eval2010.pdf

Department of Justice, Office of Justice Programs, Bureau of Justice Assistance, Center for Program Evaluation and Performance Measurement (DOJBJA). 2012. *Guide to Program Evaluation*. Washington: The Agency. URL: https://www.bja.gov/evaluation/guide/index.htm. Accessed 17 June 2012.

Employment & Training Administration. 2012. *Announcement of American Job Center Network.* Training and Employment Guidance Letter No. 36-12. Washington: The Agency. URL: http://wdr.doleta.gov/directives/attach/TEGL/TEGL_36_11.pdf. Accessed 6 August 2013.

Fraker, Thomas, Peter Baird, Alison Black Arif Mamun, Michelle Manno, John Martinez, Anu Rangarajan, Debbie Reed. 2011. The Social Security Administration's Youth Transition Demonstration Projects: Interim Report on Colorado Youth WINS. Washington, DC: Mathematica Policy Research.

Frey, William D., Robert E. Drake, Gary R. Bond, Alexander L. Miller, Howard H. Goldman, David S. Salkever, and Steven Holsenbeck. 2011. *Mental Health Treatment Study Final Report*. Rockville, Maryland: Westat, Inc.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington: The World Bank.

Government Accountability Office (GAO). 2002. *SGA Levels Appear to Affect the Work Behavior of Relatively Few Beneficiaries, but More Data Needed*. Report GAO-02-224. Washington: The Agency.

Government Accountability Office (GAO). 2008. *Social Security Disability: Management Controls Needed to Strengthen Demonstration Projects*. Report GAO-08-1053. Washington: The Agency.

Government Accountability Office (GAO). 2012. *Designing Evaluations: 2012 Revision*. GAO-12-208G. Washington: The Agency. URL: http://www.gao.gov/assets/590/588146.pdf. Accessed 17 June 2012.

Granger, Robert C. 2011a. "The Big Why: A Learning Agenda for the Scale-Up Movement." *Pathways*, Winter, 28-31. URL:

Granger, Robert C. 2011b. "Leveraging i3 to Improve Program Effectiveness." Remarks prepared for a meeting of grantees funded by the Investing in Innovation Fund (i3), Washington, D.C., January 19, 2011b.

Greenberg, David H. and Philip K. Robins. 2011. "Have Welfare-to-Work Programs Improved Over Time in Putting Welfare Recipients to Work?" *Industrial & Labor Relations Review*, 64(5), 910-948.

Haskins, Ron, and Jon Baron. 2011. "The Obama Administration's evidence-based social policy initiatives: An overview." Part 6 of National Endowment for Science, Technology and the Arts (NESTA), *Evidence for Social Policy and Practice: Perspectives on how research and evidence can influence decision making in public services*. London: NESTA, pp. 28-35. URL: www.nesta.org.uk/library/documents/Expert_Essays_webv1.pdf. Accessed 17 June 2012.

HM Treasury (United Kingdom). 2006. *The Green Book: Appraisal and Evaluation in Central Government*. London: The Agency. URL: http://www.hm-treasury.gov.uk/d/green_book_complete.pdf. Accessed 9 August 2012.

HM Treasury (United Kingdom). 2011. *The Magenta Book: Guidance for evaluation*. London: The Agency. URL: http://www.hm-treasury.gov.uk/d/magenta_book_combined.pdf. Accessed 17 June 2012.

Hoynes, Hilary W., and Robert A. Moffitt. 1999. "Tax Rates and Work Incentives in the Social Security Disability Insurance Program: Current Law and Proposed Reforms," *National Tax Journal*, 52, 623-654.

Livermore, Gina A. 2003. *Wage Reporting and Earnings-Related Overpayments in the Social Security Disability Programs Status, Implications, and Suggestions for Improvement.* Report to the Ticket to Work and Work Incentives Advisory Panel. Washington, DC: Cornell Center for Policy Research.

Liu, Su, and David Stapleton. 2010. "How Many SSDI Beneficiaries Leave the Rolls for Work? More Than You Might Think." Mathematica Policy Research Disability Policy Research Brief 10-01. Washington: Mathematica Policy Research, Inc.

Morris, Stephen, Herta Schönhofer, and Michael Wiseman. 2012. *The Design and Commissioning of Counterfactual Impact Evaluations: A Practical Guidance for ESF Managing Authorities*. Brussels: European Commission Directorate General for Employment, Social Protection, and Social Inclusion (forthcoming).

Office of Management and Budget (OMB). 2011 *Analytical Perspectives, Budget of the United States Government, Fiscal Year 2012*. Washington: The Agency. URL: http://m.whitehouse.gov/sites/default/files/omb/budget/fy2012/assets/management.pdf. Accessed 17 June 2012.

Office of Management and Budget (OMB). 2012a. *Analytical Perspectives, Budget of the United States Government, Fiscal Year 2013*. Washington: The Agency. URL: http://www.whitehouse.gov/sites/default/files/omb/budget/fy2013/assets/spec.pdf. Accessed 17 June 2012.

Office of Management and Budget (OMB). 2012b. Use of Evidence and Evaluation in the 2014 Budget. Memorandum M-12-14. URL: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-14.pdf. Accessed 13 June 2012.Patton, Michael Quinn. 2008. *Utilization-Focused Evaluation*. 4th ed. Thousand Oaks, California: Sage Publications.

Patton, Michael Quinn. 2008. *Utilization-Focused Evaluation*. 4th ed. Thousand Oaks, California: Sage Publications.

Pike, Peter, Kendra J. Alfson, and Nancy Koester. 2010. *Colorado Youth WINS Local Process Evaluation*. Denver, Colorado: Colorado WIN Partners. URL: http://www.ucdenver.edu/academics/colleges/medicalschool/departments/pediatrics/research/programs/WIN/Documents/Reports/SSA_Final_Report%203-19-2010.pdf; Accessed

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. Washington, DC: Mathematica Policy Research.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, CA: SAGE Publications.

Schimmel, Jody, Allison Roche, and Gina Livermore. 2011. *Evaluation of the Recent Experience of the Work Incentives Planning and Assistance (WIPA) Program: Beneficiaries Served, Services Provided, and Program Costs: Final Report*. Baltimore, Maryland: Social Security Administration. URL: https://www.socialsecurity.gov/disabilityresearch/documents/WIPA%20Update%20September%202011.pdf (accessed 5 May 2012).

Shipman, Stephanie and others. 2006. *Evaluation Dialogue between OMB Staff and Federal Evaluators: Digging a Bit Deeper into Evaluation Science*. Washington: Federal Evaluators. URL: http://www.fedeval.net/docs/omb2006briefing.pdf. Accessed 8 August 2012.

Social Security Administration (SSA). 2010. *Office of Program Development and Research Demonstration Project Guidebook*. Washington: The Agency. (Not publicly available.)

Social Security Administration (SSA). 2011. *Annual Statistical Report on the Social Security Disability Insurance Program, 2010*. Publication No. 13-11827. Washington: The Agency.

Stapleton, David, Stephen Bell, David Wittenburg, Brian Sokol, Debi McInnis. 2010. *BOND Final Design Report*. Baltimore, Maryland: U.S. Social Security Administration. URL: http://www.ssa.gov/disabilityresearch/documents/BOND_Design%20Report_FINAL_Del2-2_12-17-10.pdf.

Weathers, Robert R., and Jeffrey 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.