

# **Design Options for an Evaluation of TANF Job Search Assistance**

**OPRE Report No. 2013-01**

**February 2013**

# Design Options for an Evaluation of TANF Job Search Assistance

**OPRE Report No. 2013-01**

**February 2013**

Laura R. Peck, Stephen H. Bell, Jacob Alex Klerman, and Randall Juras

Submitted to:

Erica Zielewski and Seth Chamberlain, Project Officers  
Office of Planning, Research, and Evaluation  
Administration for Children and Families  
U.S. Department of Health and Human Services

Contract Number HHSP23320095624WC

Project Director: Robin Koralek  
Abt Associates Inc.  
4550 Montgomery Avenue, Suite 800N  
Bethesda, MD 20814

This report is in the public domain. Permission to reproduce is not necessary. Suggested citation:  
Peck, Laura R., Stephen H. Bell, Jacob Alex Klerman, and Randall Juras (2013). *Design Options for an Evaluation of TANF Job Search Assistance*. OPRE Report # 2013-01, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

## Disclaimer

The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at <http://www.acf.hhs.gov/programs/opre/index.html>.



## Table of Contents

|  |           |
|--|-----------|
| <b>Overview .....</b>  | <b>1</b>  |
| <b>1. Introduction .....</b>   | <b>2</b>  |
| 1.1 Study Overview and Purpose of this Report .....  | 2         |
| 1.2 Guiding Principles for Evaluation Design .....   | 2         |
| 1.3 The Design Challenge .....   | 3         |
| 1.4 Plan for the Report .....  | 5         |
| <b>2. Define the Interventions to be Tested and Research Questions .....</b>                   | <b>6</b>  |
| 2.1 What Do We Mean by the Job Search Process and Job Search Assistance? .....                 | 6         |
| 2.2 How Do JSA Programs Achieve their Policy Goals? .....                                      | 8         |
| 2.2.1 Labor Market Outcome #1—Time Spent Searching for Employment .....                        | 10        |
| 2.2.2 Time Spent Working in a Job—Job Quality .....  | 11        |
| 2.3 What Aspects of JSA Programs Should We Evaluate? .....                                     | 12        |
| 2.4 What are the Appropriate Research Questions? .....   | 14        |
| 2.5 Discussion .....   | 18        |
| <b>3. Define Key Measures and Possible Data Sources .....</b>                                  | <b>19</b> |
| 3.1 Individual Context .....   | 20        |
| 3.2 Outputs and Similar Services Outside the System .....                                      | 22        |
| 3.3 Outcomes—Short-Term and Long-Term .....  | 24        |
| 3.4 Multiple Comparisons and Confirmatory Outcomes .....                                       | 27        |
| 3.5 Discussion .....   | 28        |
| <b>4. Design Options for Measuring Causal Impacts on Individuals .....</b>                     | <b>30</b> |
| 4.1 External Validity .....  | 30        |
| 4.1.1 The Goal and the Challenge .....   | 30        |
| 4.1.2 Design Responses: Choosing or Approximating a Representative Sample .....                | 31        |
| 4.1.3 Analysis Responses: Enhancing External Validity with a Nonrepresentative<br>Sample ..... | 32        |
| 4.2 Multisite Considerations .....   | 32        |
| 4.3 Internal Validity .....  | 34        |
| 4.4 Individual-Level Randomized Designs .....  | 35        |
| 4.4.1 Basic Randomized Experimental Design .....   | 35        |
| 4.4.2 Multi-Arm Randomized Experimental Design .....   | 36        |
| 4.4.3 Randomized Factorial Design .....  | 38        |
| 4.5 Cluster-Randomized Designs .....   | 40        |
| 4.5.1 Randomized Sites (Without Individual Within-Site Random Assignment) .....                | 41        |
| 4.5.2 Randomized Sites Plus Individual Random Assignment to Components within<br>Sites .....   | 42        |
| 4.5.3 Methods for Estimating Impacts .....   | 43        |
| 4.6 Additional Analytic Methods .....  | 47        |

|           |   |           |
|-----------|---|-----------|
| 4.7       | Conclusion .....  | 49        |
| <b>5.</b> | <b>Design Options for Measuring General Equilibrium Effects .....</b>   | <b>50</b> |
| 5.1       | General Equilibrium Effects in a Job Search Context.....  | 50        |
| 5.2       | Possible Approaches to Estimating General Equilibrium Effects.....  | 51        |
| 5.2.1     | Cluster-Randomized Designs.....   | 52        |
| 5.2.2     | Exploiting LLM-Level Nonrandom Variation.....   | 53        |
| 5.2.3     | Double-Randomized Designs.....  | 54        |
| 5.2.4     | Utilizing Nonintervention Local Labor Markets to Estimate Displacement.....   | 54        |
| 5.2.5     | Job Search Models Using Experimental Data Only.....   | 55        |
| 5.3       | Implications for DOSE.....  | 56        |
| 5.4       | Obtaining a General Equilibrium Impact Estimate from within the Basic Experiment.....   | 57        |
| <b>6.</b> | <b>Determine Needed Sample Sizes .....</b>  | <b>58</b> |
| 6.1       | Individual-Level Randomized Options.....  | 58        |
| 6.2       | Cluster-Randomized Options.....   | 63        |
| 6.3       | General Equilibrium Effect Analysis.....  | 65        |
| 6.4       | Data Collection Strategies Revisited.....   | 66        |
| 6.5       | Discussion and Conclusion.....  | 68        |
| <b>7.</b> | <b>Evaluation Components to Complement a JSA Impact Study .....</b>   | <b>70</b> |
| 7.1       | Process Analysis .....  | 70        |
| 7.2       | Participation Analysis .....  | 72        |
| 7.3       | Benefit-Cost Analysis .....   | 74        |
| 7.4       | Recommendations for Expanding the Design and Implementation Effort .....  | 76        |
| <b>8.</b> | <b>Conclusion.....</b>  | <b>78</b> |
| 8.1       | The Guiding Principles Revisited: Assessing the Tradeoffs Across Varied Design Options.....   | 79        |
| 8.1.1     | Research Merit.....   | 79        |
| 8.1.2     | Technical Merit.....  | 79        |
| 8.1.3     | Practical Considerations.....   | 81        |
| 8.1.4     | Policy Relevance.....   | 83        |
| 8.2       | Conclusion .....  | 85        |
|           | <b>References.....</b>  | <b>86</b> |
|           | <b>Appendix A: Bias in Experimental Impact Measures when JSA Participants Displace Other Workers: Schematic Analysis and Possible Solutions within the Experiment .....</b> | <b>90</b> |

## Overview

This report presents the results of a 15-month evaluation design effort funded by the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation to develop options for a rigorous evaluation of the impact of Job Search Assistance (JSA) services on low-income workers. A partner report reviews current JSA programs and evidence on their effectiveness.

The report presents a logic model as a foundation for specifying research questions for an evaluation. Though many important research questions exist, the report recommends focusing on the differential impact of JSA program variants on short-term earnings. Learning about the relative effectiveness of program components has the potential to inform program improvement and efficiencies, and the report argues that short-term earnings is the most policy relevant outcome.

The core of the report describes several broad designs for estimating causal impacts of JSA on individual job seekers, using both individual- and group-randomized experimental designs. It also discusses analytic methods and issues of the generalizability of impact findings over time, geographic areas, and participant characteristics. For various designs, the report estimates the sample sizes required to detect impacts of an important policy magnitude. The report notes that quarterly earnings are recorded in administrative data (state Unemployment Insurance for example), while other key outcomes—such as time to employment, hourly wages, household income and poverty—are not. The implication is that using administrative data would allow for lower evaluation costs, especially where large samples would be needed.

In addition, the report discusses how implementation, participation, and benefit-cost analyses might complement and inform impact analyses. It also discusses general equilibrium effects that exist in the labor market when workers (and firms), beyond direct JSA recipients, are affected by JSA interventions. It suggests analyses for estimating the direction and magnitude of these effects.

The report concludes with a discussion of the tradeoffs across various evaluation design options when the criteria of research, technical merit, practical merit, and policy relevance are applied. Specifically, the conclusion explains that JSA is a low-intensity intervention, which is likely to have only small effects. The differential impacts of JSA program variants will be smaller yet, with large samples needed to detect them. In response, the report suggests a “maximum variation” comparison, creating a contrast between treatment options that might generate impacts of greater magnitude than a straight comparison of various versions of JSA. Comparing a bare minimum JSA program with a services-rich JSA program would likely generate larger estimated impacts (relative to testing the effectiveness of a single component), reducing sample size needs and therefore evaluation costs and logistical demands. Given that states currently invest resources in JSA programs, despite it being unclear how much these resources help participants, this type of evaluation design has the potential to provide useful information for policy design and decisions. For example, if no difference is detected between bare minimum and services-rich JSA, then states could shift resources from JSA to other efforts. Conversely, if an impact larger than the cost of the program is detected, then states can justify their investments in JSA, knowing that it makes a difference.

## 1. Introduction

### 1.1 Study Overview and Purpose of this Report

To address the need for more recent and relevant research on the effectiveness of job search assistance (JSA) provided to disadvantaged workers and heads of households, the U.S. Department of Health and Human Services (HHS), Administration for Children and Families (ACF) Office of Planning, Research and Evaluation (OPRE) awarded a contract for the Design Options of the Search for Employment (DOSE) study to Abt Associates Inc. in September 2011. The purpose of this contract was to inform ACF's thinking on JSA approaches and potential strategies to evaluate them. As part of this contract, Abt produced *Job Search Assistance Programs: A Review of the Literature* (hereafter Klerman et al., 2012), which describes existing JSA programs and reviews the literature on their impacts. Building on that effort, this document explores options for the design of an evaluation of the impact of JSA programs on those receiving public assistance, to help ACF move forward with such an evaluation in a potential future procurement. The approach outlined in this document represents Abt's views on the contours of a potential evaluation of JSA approaches and does not necessarily reflect ACF's future research agenda or the design of a future JSA evaluation.

As noted in Klerman et al. (2012), for most Americans, employment is critical: jobs provide economic stability and work anchors the day. Usually, employment is preceded by some form of job search; often, searching for a better job continues after an initial job is found. Effective job search methods are therefore of great importance. JSA programs—short-term, relatively low-intensity, relatively low-cost programs to help job seekers find jobs—are also a key component of many government-funded assistance programs and are available to workers generally. These JSA programs assist a job seeker both in (1) achieving her personal goal of getting employed and (2)—as a *quid pro quo* of receiving economic assistance while out of work—encouraging more intensive search leading to more job offers and quicker acceptance than might otherwise be the case.

Despite the crucial role of job search in the lives of families and of JSA in supporting income assistance programs, little recent research considers the relative effectiveness of various job search methods and of the components of JSA programs. Going back several decades, a moderate amount of research exists on job search assistance bundled with other services such as vocational training, but much less research has evaluated the distinct contributions of individual components of the bundle. While existing research has considered the impacts of JSA programs on broad populations of job seekers, less research has considered how those impacts vary across subgroups defined by demographic characteristics and past employment and education. Some of this research is focused on American welfare recipients, but much of it is focused on American Unemployment Insurance (UI) program participants or on Europeans. Finally, much of this literature is now several decades old, and the labor market—in general, and for disadvantaged workers in particular—has changed, as has the public policy context. An opportunity exists to provide fresh evidence on the impacts of job search assistance, and such evidence would be welcome in both scholarly and policy circles.

### 1.2 Guiding Principles for Evaluation Design

The goal of this document is to explore possible designs for a rigorous impact evaluation of JSA in order to provide new evidence on the effectiveness of JSA programs. Rather than designing some new JSA program, the project explores the impact of JSA program service provision modalities and programmatic

components that are in relatively broad use within existing Temporary Assistance for Needy Families (TANF) programs. Should an evaluation be undertaken, ACF will decide which particular intervention(s) in this domain to evaluate. The current report seeks to provide useful guidance for evaluation design regardless of these decisions, and a generic “blueprint” for conducting such an evaluation.

In considering various evaluation designs, we propose the following “guiding principles” for ACF to consider in picking a JSA evaluation agenda for the future, and utilize these principles in developing our design recommendations:

- **Research merit.** An evaluation design’s ability to address the right question(s) is an essential criterion. In other words, research merit considers the extent to which the research findings would fill a gap in existing knowledge. We also seek to address the right impact question for the right population, such that the external validity, or generalizability, of the evaluation design becomes part of the assessment of research merit.
- **Technical merit.** For the right impact question, the planned evaluation needs to give a scientifically valid answer. Factors that weigh into this technical merit criterion for evaluation design include accurate causal attribution of outcomes to JSA inputs, adequate sample sizes for statistical reliability, reliable data collection, and data analysis methods that address the distinctive challenges of a JSA evaluation, such as the complementarity of impacts on individuals and broader labor market general equilibrium effects.
- **Practical considerations.** Here we judge whether a particular evaluation design is feasible. Will the interventions tested be cost neutral from a program perspective (i.e., will not require additional program funds)? Do suitable and sufficient sites exist for fielding a test of the desired intervention(s)? Are a sufficient number of state and local JSA agencies in the sites likely to be willing to cooperate with an evaluation, and to vary their intervention in order to create a contrast that would allow learning about the effectiveness of specific JSA modes of service provision and/or program components? Do those agencies have the administrative capacity to manage both the programmatic and evaluation portions of a field test? Given the required sample sizes, can the necessary data for the design be collected, and its analysis reported, at a reasonable resource cost?
- **Policy relevance of the tested intervention(s).** It matters enormously that the selected field test be relevant to the public policy decisions to be informed by the research, an issue that we revisit throughout this document. We anticipate that ACF’s choice of a JSA intervention or interventions to be tested will be driven by the agency’s need/interest in informing future policy decisions.

Using these principles, the concluding chapter of the report analyzes the tradeoffs of various elements of the design options and approaches.

### 1.3 The Design Challenge

In some ways, designing an evaluation of multiple components of a JSA program draws on generic impact evaluation ideas. Considerations of internal validity suggest use of a randomized experimental design whenever feasible. Considerations of statistical precision suggest estimation using regression, rather than simple comparison of means. We compute the statistical power of various designs and urge sample sizes large enough to confidently detect effects of a relatively small magnitude, as would be expected in a test assessing variations in JSA program elements. However, in five crucial and related ways, an evaluation of multiple components of a JSA program differs from the conventional impact

evaluation problem. Much of this document is focused on those differences and related optimal design options.

First and foremost, evaluation of the *modes of service delivery* and/or *components* of a JSA program demands that we think about how best to detect *differential* impacts that are likely to be small. Differential effects of two or more variants of an already relatively light-touch intervention cannot be large in magnitude. TANF rules and TANF experience imply that JSA programs are short and relatively low intensity. We would therefore expect their overall impact to be small relative to longer and more intensive employment training programs, with differential impacts between program variants being even smaller. The operative policy question is not JSA versus no JSA; instead, the operative question is one mode of JSA versus another form of JSA. Klerman et al.'s (2012) scan of existing JSA programs identified a relatively narrow range of what constitutes JSA. A future evaluation could focus on forms of JSA that are already in widespread use, aiming to understand the variation in impacts across variations in program design. With differential impacts expected to be small in magnitude, required sample sizes (as estimated in Chapter 6) are very roughly 10 times as large as single-treatment samples in earlier ACF studies such as the National Evaluation of Welfare to Work Strategies (NEWS) or the Employment Retention and Advancement (ERA) project.<sup>1</sup>

Second, we are interested not only in these differential impacts, but also in how impacts vary with individual characteristics (e.g., education levels or prior work experience) and local context (e.g., the state of the economy, TANF sanction policy). As samples narrow to focus on subsets of individuals or places, the needed overall sample size increases to support detecting a minimum detectable (differential) impact of policy relevance.

Third, as noted above, our interest is in varying *multiple* components of JSA programs. Ideally, we would not estimate the impact of varying a single component, but instead, the impact of varying multiple components. Conventional multi-arm studies require large sample sizes—beyond the large sample sizes required to estimate small impacts and differential impacts. In Chapter 4, we explore the randomized factorial designs that support answering a greater number of research questions without substantially greater samples. That said, additional questions imply additional hypothesis tests, which means added considerations regarding statistical corrections for such multiple comparisons. Despite the design appeal, operationally, implementing an experimental factorial design in the field is likely to be challenging.

Fourth, likely large sample sizes imply a multisite study, for estimating both overall impacts and differential impacts. Consideration of contextual variation implies a site “selection” process that would deliberately choose study sites according to contextual factors of interest. In brief, design and estimation need to explicitly consider multisite issues, which may be challenging in an experimental context or

---

<sup>1</sup> NEWS had seven sites—three running a two-arm (treatment/control) study and four running a three-arm study (in three Labor Force Attachment/Human Capital Development/Control; in one Traditional/Integrated/Control), for a total of 18 arms (see Hamilton et al., 2001, Table 2.2, p. 38). Across all sites, the full impact sample was 41,715, or 2,317 per arm, or about 5,000 observations for a conventional two-arm study. A study looking for an impact slightly less than a third of that size might want a sample 10 times (i.e., slightly more than three squared; sample size goes up as the square of the inverse of the impact to be detected) as large, i.e., 50,000 observations for a two-arm study.

unsatisfying in a nonexperimental context. We will further explore the practical implications of a multisite evaluation proposition.

Fifth, even if enough sites can be recruited and individual study participants available, data collection often drives study cost. Survey data collection on large samples, such as those needed to detect small differential impacts, would, at best, border on cost infeasible. To explore important differential impact questions, we would need to consider less expensive data collection strategies. As we discuss in Chapter 3, for some—arguably the focal—outcomes, administrative data are attractive, although their use involves tradeoffs. Careful consideration of those tradeoffs, and perhaps a mixed strategy—administrative data on the full sample, a survey on a subsample—seems appropriate.

## **1.4 Plan for the Report**

To respond to these and other challenges in designing an evaluation, the remainder of this report proceeds as follows. Chapter 2 develops a logic model for JSA programs and uses that logic model to specify research questions for the evaluation. Chapter 3 then identifies the key data measures needed for the research and discusses strategies for collecting information about them. In Chapter 4, we describe several broad designs for estimating causal impacts of JSA on individual participants, using both individual- and group-randomized experimental designs. This chapter also discusses how the resulting data would be analyzed and the generalizability of the impact findings over time, geographic areas, and participant population characteristics. Chapter 5 considers general equilibrium effects in the labor market when workers (and firms) beyond direct JSA recipients are affected by the JSA intervention, and the analyses that one would undertake for estimating the direction and magnitude of these broader effects. For various designs, Chapter 6 considers the sample sizes required to detect impacts of an important policy magnitude. Chapter 7 then discusses the possible role of implementation, participation, and/or benefit-cost analyses to complement and inform the impact analyses. Finally, Chapter 8 concludes with a discussion of the tradeoffs across various design options when the criteria of research, technical, and practical merit are applied to the methodologies developed in earlier chapters. It is our hope that ACF can use this assessment to launch a valuable, highly informative evaluation of the impact of JSA services provided to public assistance populations in today's world and in the future.

## 2. Define the Interventions to be Tested and Research Questions

This chapter addresses four questions: What do we mean by the job search process and job search assistance (JSA) programs (Section 2.1)? How do JSA programs achieve their policy goals (Section 2.2)? What aspects of JSA programs should we evaluate (Section 2.3)? And, what are the appropriate research questions in such an evaluation (Section 2.4)? The chapter concludes with a forward-looking discussion of each of these questions to consider implications for evaluation design and the rest of the report (Section 2.5).

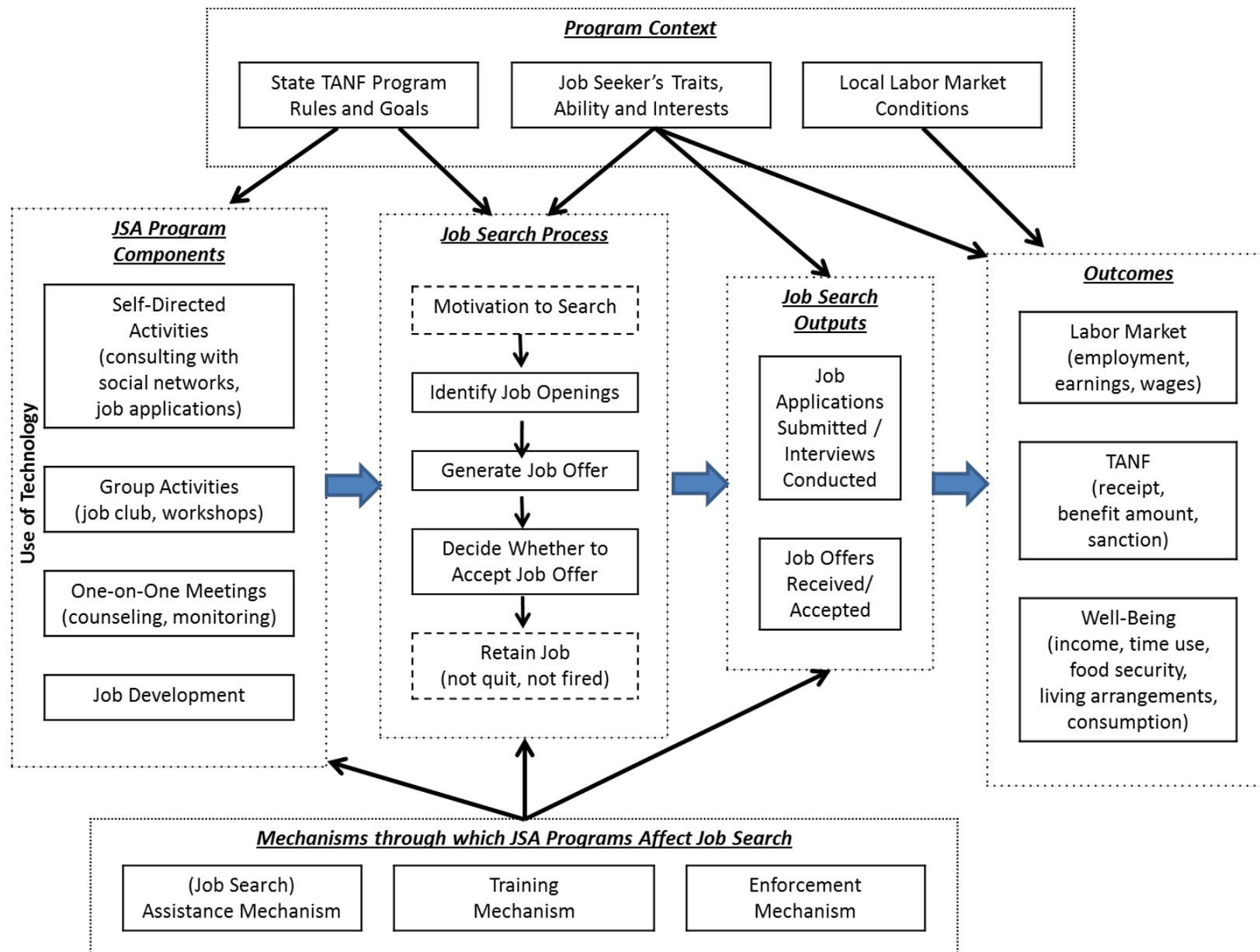
### 2.1 What Do We Mean by the Job Search Process and Job Search Assistance?

To be employed, an employee has to seek out, find, and obtain a job. Even after being offered a job and holding it, job search often continues. Searching for and obtaining a “better” job is used by job seekers as a strategy for labor market advancement—both on their own and with the help of others.

Federal and state governments provide support through several JSA programs. Government-supported job search activities tend to be short (a few weeks), relatively low-intensity, and relatively low-cost programs (Klerman et al., 2012). Additionally, a large body of research has studied the general effects of job search assistance. However, despite evidence on their general effectiveness, very little is known about the comparative effectiveness of alternative modalities of job search and the marginal effects of independent components of job search assistance. The aim of this current design study is to develop evaluation options that would fill these gaps.

In designing an evaluation of job search and JSA programs, we begin with the conceptual framework of job search proposed in our earlier report (see Exhibit 2.1 and Klerman et al., 2012). For the purposes of this evaluation design, the importance of this framework—illustrated graphically in that report—is that job search activities can be conceived of as a set of modalities composed of a series of component parts, which are expected to influence the individual search process. Based on the success of this search process, job search produces a set of outputs, such as applications submitted, interviews completed, and job offers received and accepted. These outputs lead to employment and earnings as the primary outcomes, as well as secondary outcomes including reductions in public assistance, more consumption, but less leisure—with implications for overall well-being. In its simplest form, an evaluation of JSA programs would be designed to yield a more complete and/or nuanced understanding of the causal relationship between job search inputs (modalities, components and stages in the search process) and search outputs and outcomes.

Exhibit 2.1: A Conceptual Framework for JSA



In addition to producing deeper insight into the internal elements that influence search outcomes, a strong evaluation might also produce an understanding of the contextual factors that moderate the effects of a program, and the policy mechanisms that serve as the particular levers that mediate impacts. Often times policymakers, program administrators, and researchers want answers to questions such as “What job search activities work best in which labor market environments?”; or “What job search approaches work best for certain subgroups of job seekers?”; or “What policy mechanisms are most likely to increase earnings or reduce public assistance?” Unfortunately, few evaluations are designed to provide answers to these questions and, despite being a widely covered topic, there is little systematic evidence on these elements of the job search process. Our framework incorporates these elements, and we will aim to understand them as rigorously as possible when designing the evaluation.

In an ideal world, an evaluation of job search programs consistent with this framework would have sufficient variation in modalities, components, contexts, stages in the process, mechanisms, outputs, and outcomes such that it would produce experimental estimates of the independent effects of each contributing factor in this causal process.<sup>2</sup> Were such an evaluation possible, it would answer two core research questions, each of which embeds several followup questions, which are the essence of this research:

1. What are the relative impacts of selected JSA program traits on each output and outcome? In addition, what specific modalities and components work best for which groups of job seekers?
2. How do contextual factors (including local labor market, policy setting, and individual characteristics) moderate effects? From the perspectives of the individual, the government, and society as a whole, what are the costs and benefits of variation in the configuration of JSA?

The remainder of this chapter explores a series of important factors that must be addressed even before evaluation options can be developed.

## 2.2 How Do JSA Programs Achieve their Policy Goals?

Before we can consider evaluating the effectiveness of any JSA program, we must first define the inputs, outputs, and outcomes to be measured. In Klerman et al. (2012), we have defined the modalities and components of JSA programs. We do not repeat that discussion here, but rather incorporate and extend it to examine outputs and outcomes. We have also previously discussed outputs and potential outcomes, but not in the context of designing an evaluation.

The right side of our conceptual framework portrays two sets of job search results flowing sequentially from the job search process. First are *outputs* of the job search process. The figure includes four primary outputs of a job search process: job applications submitted, interviews completed, offers received, and

---

<sup>2</sup> This discussion deliberately does not include a “no job search” option. Because an evaluation of JSA would be designed principally to address job search in the context of the TANF program, not seeking employment or not being employed is not an option. Moreover, a “no job search” option is neither realistic nor helpful to policymakers and program administrators. Given the role of search for employment in the lives of low-skilled and low-wage workers, many social programs include a JSA component. Totally denying JSA seems sufficiently unlikely as to not be worth explicit evaluation. Instead, the important policy questions consider how to structure JSA programs, i.e., JSA program model A versus JSA program model B (not versus no JSA at all).

offers accepted. The language of outputs implies a quantitative dimension to these factors—counting the number of applications submitted, the number of interviews completed, and the number of offers received. Implicitly, a qualitative dimension of job applications and interviews also exists: to what extent do participants have the job search skills to identify appropriate jobs for which to apply, complete applications correctly, and perform well in interviews? Further, to what extent do they put forth a sincere effort in doing so? Three job-seeking scenarios illustrate the complexity of the relationship between the quantitative and qualitative dimensions of job search outputs.

Imagine a scenario in which a strong candidate is seeking employment. She may only need to submit one application, complete one interview, receive one offer, and take that one job. Alternatively, because of her advantage in the market, she may choose to submit many applications, explore options through a number of interviews, receive multiple offers, and not immediately take the first one, but instead continue her search and leverage the multiple offers to her advantage. Similarly, a weak candidate may submit just as many applications, go on just as many (if not more) interviews, and receive and accept the only offer that comes to him. In these three scenarios, varying quantitative and qualitative elements illustrate how job search outputs ultimately converge on one crucial output—job offers—one of which is eventually accepted. Specific measures of these elements of the job search process and its outputs would do well to account for both the quantitative and qualitative characteristics of the experience.

On the far right side of the framework are three categories of *outcomes* which the job search process is intended to affect:

- **Labor market outcomes** including employment or unemployment spells and duration (e.g., time to first job, duration of that job, time to subsequent jobs, and length of those jobs), earnings, hourly wages, and other aspects of the accepted job (e.g., benefits, number and regularity of hours, shift, and regularity of shift).
- Increased employment and earnings should affect the *use of public assistance*—namely, reducing TANF benefit payments and, ideally, increasing exits from TANF. However, increased oversight and/or enforcement of a requirement to search might also result in additional sanctions. Thus, TANF program outcomes including receipt of any TANF cash benefit, the amount of the benefit, and unusually high or low sanction rates are key indicators of the success of JSA.
- Ultimately, the goal of TANF and its activities, including JSA, is to help disadvantaged families live a better life economically and in other ways. Hence, JSA programs seek to affect **broader well-being**, including total family income, time use, food security, consumption, living arrangements, family stability, child health, child behavior, and child academic achievement. Beyond labor market outcomes, these broader well-being outcomes are important because earlier studies of welfare-to-work programs suggested that any increase in earnings was offset by the loss of means-tested benefits (such as TANF and Supplemental Nutrition Assistance Program [SNAP], even considering the Earned Income Tax Credit [EITC]) and less time for child rearing, home making, and pure leisure (e.g., Hamilton et al., 2001).

The primary area of interest is labor market outcomes, for a number of reasons. First, while it is indeed the case that many job search assistance programs and components include elements that are designed to reduce the use of public assistance (such as the use of sanctions for noncompliance) and to improve other factors associated with well-being (such as the psyche and resiliency of job seekers), what distinguishes job search assistance programs is their primary focus on employment and a job as the means to achieving

these other ends. Second, and conversely, for the purposes of this evaluation, job search assistance interventions are defined such that employment is their primary outcome. Contrast this with a program of home visiting for new mothers or with a program solely focused on teaching basic skills. Both of these programs may have effects on employment, public assistance, and broader well-being (including child outcomes), and may do so by altering individuals' job search processes; however, they are primarily designed to serve parenting or educational purposes. As first-order outcomes, and to test program theory, labor market outcomes should be the primary outcomes of interest in a JSA evaluation.

Settling on labor market outcomes as a principle focus, however, does not resolve the discussion. In fact, it narrows it to perhaps the most important aspect of understanding the impacts of JSA programs—the tradeoff between time spent searching for a job and time spent working in a better/worse job.

### 2.2.1 Labor Market Outcome #1—Time Spent Searching for Employment

One possibility is that JSA programs accelerate employment, shortening the time to securing a job. This might occur through multiple pathways, including:

- Pointing the job seeker to a specific job opening (e.g., the job development modality).
- Teaching job search skills such that the job seeker finds a job on her own faster (i.e., the assistance mechanism).
- Teaching world-of-work skills such that the participant presents better in a job interview (again, the assistance mechanism).
- Increasing motivation to work (e.g., assistance mechanism efforts to build self-esteem, or enforcement mechanism monitoring of job search efforts and threats of sanction).

For modalities and components that deliberately focus on this pathway to generating change, reducing the time to the next job implies no effect on future employment outcomes beyond those that stem from the timing of that next job. Or, put another way, in the rapid employment model, the strategies for advancement come not from holding out for a “better” job, but rather through being attached to the workforce, by being employed, and by using existing employment as a stepping stone.

This raises two important implications for our evaluation design. First, considering time to the next job suggests a focus on very proximate, relatively short-term outcomes (perhaps as short as a few weeks or months after entry into the JSA program). This focus is embedded in the current programmatic structure of TANF. For the purposes of inclusion in the TANF state work participation rate, job search and job readiness activities are limited to no more than four consecutive weeks and no more than six weeks over a 12-month period. However, 12 weeks are allowed in states that meet the conditions of a “needy state,” as defined for Contingency Fund purposes. TANF rules allow states to calculate participation on an hourly, rather than weekly basis, with one week of participation defined as 30 hours for work-eligible individuals (20 hours for single parents with children younger than six). Thus, if the JSA program is to lead to a job, it will almost always do so within that time window. Conversely, if there are no impacts within this short time window, effects over a longer time window are unlikely.<sup>3</sup>

---

<sup>3</sup> The implication of “if no short-term impacts, then no long-term impacts” is plausible for true impacts. That said, as we discuss later (in Chapter 3), longer followup can increase precision of impact estimates. A short-

Second, these programs and their underlying theory imply that the immediate focus on labor market outcomes should be related to duration and intensity of attachment and participation in the labor market rather than hourly wage levels or earnings. An illustrative example is helpful here.

Imagine two job seekers, Harry and Alice.

- At the beginning of the quarter, Harry starts a rapid employment job search program—the Right Now Works program—and after two weeks in the program finds a job that pays him \$10 per hour, working 35 hours per week. For the quarter, **Harry earns \$3,850** ( $\$10/\text{hour} \times 35 \text{ hours/week} \times 11 \text{ weeks of work}$ ).
- Alice also starts a job search program at the beginning of the quarter but she is enrolled in a job search program that focuses on finding better jobs and better job matches between employees and employers—The Right Job Works program. Alice spends four weeks searching for work and finds one that pays her \$11 per hour for 40 hours of work per week. For the quarter, **Alice earns \$3,960** ( $\$11/\text{hour} \times 40 \text{ hours/week} \times 9 \text{ weeks of work}$ ).

At the end of one quarter, if the outcome of interest is hourly wages or earnings, Alice is doing better than Harry. However, both programs did what they were designed to do, raising the question of whether this is a reasonable comparison. As we later note, the evaluation’s research questions focus on these varied labor market outcomes—including time to employment, wages, and hours—but, for reasons we discuss in the next chapter, we choose overall earnings from work as the primary, confirmatory outcome. As this example shows, the earnings measure is a useful summary measure, capturing existing variability. Additional exploratory analyses can tease apart the underlying variability in experience across multiple measures, including job quality, to which we turn next.

### 2.2.2 Time Spent Working in a Job—Job Quality

The example of Harry and Alice discussed above illustrates how elaborated program theory might suggest impacts on other aspects of the labor market outcomes. For example, given that jobs are heterogeneous, JSA programs might change the jobs obtained:

- **A better job match:** JSA programs improve job search skills and (possibly) improve job matches, i.e., helping job seekers to find “better” jobs, which can increase wages and earnings. Similarly, better job matches resulting in a job that the seeker likes more may cause the job to last longer. Given that short job tenure is common in the low-skill labor market, impacts on duration of the job are possible and could have substantial impacts on quarterly earnings (even in the absence of an impact on wages). While enhanced soft skills may result in an employer who is more satisfied with the job seeker, now worker, and also lead to longer stints of employment, soft skills and other activities that work through the “training mechanism” are excluded from this report’s exploration of design options.

---

term impact is possible though might not be detectable for a given sample size, whereas a longer-term impact (pooling over a longer time) would be detectable for that same sample size.

- **A worse job match:** Conversely, the enforcement mechanism might cause job seekers not to search long enough. This might lead to poorer job matches, lower wages, and shorter job durations; and taken together, lower earnings.

While the previous discussion calls out three key dimensions of a job match (wage rate, duration, and earnings), other dimensions of a job match are also important, including fringe benefits provided (e.g., health insurance, paid vacation and sick leave, and retirement benefits), hours of work (number, shift, and regularity), and other conditions of employment (e.g., occupation, workplace safety, positive work environment). To some extent, these dimensions of a job match are matters of fit. Some people prefer indoor work; others prefer outdoor work. Some would prefer more cash to more benefits; others would prefer the opposite. Sometimes through job search assistance, the job seeker can get more of both; or perhaps less of both, if induced to take a job too soon.

Furthermore, while it is most plausible that any impacts will occur in six months to a year after enrollment in the JSA program, longer-term impacts are possible. A good job match might last a long time and provide experience and skills that lead to growing wages and earnings—in this job and in future jobs. The discussion in Klerman et al. (2012) suggested that such impacts are likely to be small, but they are possible, and there is some possibility that they might be large.

Moreover, previous research from the evaluation of Job Training Partnership Act programs, the NEWWS, and the ERA project found that job loss is quite common among low-skilled, low-wage workers (even when there are higher wages and “better” matches). Results from those studies can be read as suggesting that programs that focus on maintaining a connection to the labor market are more likely to increase total earnings over many years.

Finally, any program that affects labor market outcomes will likely affect other outcomes related to receipt of benefits from public programs—both mechanistically (i.e., through benefit formulas) and behaviorally (i.e., through decisions to forgo small benefits to which one might be entitled—due to some combination of stigma, pride, and cost of time relative to the size of the benefit). As a result, we might expect to see impacts on TANF benefits paid, as well as SNAP benefits received, EITC benefits received, and Medicaid enrollment.

In sum, JSA programs are expected to primarily affect labor market outcomes—either labor force participation or hourly wages and earnings. In addition, from their impact on these outcomes, we expect that, in a TANF context, JSA may also have impacts on public assistance and other measures of well-being (including child outcomes). However, modalities and components of JSA approaches can differ in design and implementation, and be used at different intensities, and this can have an effect on outcomes. The evaluation design and data collection strategies and measures must take these elements into consideration in addition to the theory of change that conceptually guides JSA policy and program design.

### 2.3 What Aspects of JSA Programs Should We Evaluate?

The conceptual framework depicted in Exhibit 2.1 suggests several dimensions of JSA programs that might be evaluated. An evaluation of JSA could vary basic intervention *mechanisms* (bottom row of the exhibit), service delivery *modalities*, or elemental job search *components* within modality or mechanism (both in the far left column of Exhibit 2.1). With respect to mechanisms, an intervention might aim to have an impact through a specific mechanism, or through multiple mechanisms. With respect to modalities, an intervention might specify the allocation of JSA program staff between one-on-one-

activities (e.g., case management), group activities (e.g., job club), and activities not directly involving participants (e.g., job development). In terms of components, an intervention might specify whether to include, or the intensity of, one or a group of activities in one or more of the steps of the participant's job search process (the second column of Exhibit 2.1)—motivation to search, identifying job openings, generating job offers, and job retention.

Klerman et al. (2012) reviewed the existing literature on the impacts of JSA program components aimed at various steps of the job search process. The literature is thin; much of it is old and focused on UI recipients (rather than Aid to Families with Dependent Children [AFDC] or TANF recipients), and there are few clear findings with respect to components of JSA programs acting on steps of the job search process.

It seems likely that impacts of interventions will vary with the characteristics of the TANF programs in which they are situated. A program that has large potential sanctions, strong oversight of participation, and swift sanctioning of nonparticipation is likely to see less benefit from additional efforts towards the enforcement mechanism, for example. Or perhaps conversely, if sanctioning is weak in general, then oversight of job search might be ineffective: the punishment for nonparticipation is low and oversight in the post-JSA program phase will be weak. Perhaps in programs with high benefits there is reason to stay on TANF, such that enforcement would matter; but in programs with low benefits there is already an incentive to find a job as fast as possible. Such arguments suggest that impacts may vary with the details of the broader TANF program. From a design perspective, this suggests testing the intervention in a range of broader TANF program settings.

The state of knowledge is slightly different with respect to mechanisms. Here, the literature is more informative. While the evidence is not conclusive, drawing on studies across the U.S. and Europe, and in welfare programs as well as in unemployment insurance programs, Klerman et al. (2012) argue that evidence of impact of low-intensity efforts through the assistance mechanism or through the training mechanism is at best mixed. In contrast, stronger evidence exists from funding streams other than U.S. TANF programs (e.g., U.S. UI programs, European unemployment and welfare programs) that the enforcement mechanism—i.e., additional oversight of job search and sanctioning of those who do not search actively—leads to fewer people receiving benefits, smaller benefit payments, and shorter job search (at least over a horizon of a few months). Evidence of an impact of the enforcement mechanism on earnings is more mixed.

On what JSA program configuration(s) to explore, ACF has given guidance to focus on aspects of JSA programs that would work through the assistance mechanism or through the enforcement mechanism. ACF also stated that intensive efforts through the training mechanism are out of scope for this project because they involve hard skills training and not solely JSA. Consistent with this guidance, this document focuses on evaluation of efforts that work through the assistance and enforcement mechanisms. We therefore consider variation in JSA service provision modalities and JSA components that relate only to these two mechanisms. That said, we recognize that any evaluation focused on the assistance and enforcement mechanisms will operate in a world in which variation in training mechanism provision exists; and site selection will need to consider that sources of variation—as well as other stated sources—to ensure that results are not confounded accordingly.

Beyond this guidance about mechanisms, ACF has indicated a desire to keep options open as to what combination of service delivery modalities and specific components of JSA programs targeted at which steps of the job search process should be evaluated. In particular, ACF has indicated a desire to test what

is currently in use in some JSA programs, acknowledging that more complete identification of JSA program configurations that are currently in use would require a broader field effort to consult experts and practitioners than was undertaken in the current study.

## 2.4 What are the Appropriate Research Questions?

In this section, building on discussions with ACF, we propose 12 research questions (RQs) to be considered for inclusion in a study of the impact of JSA programs for TANF recipients. (Questions related to implementation, program participation, and cost benefit analysis, which are also important to understanding JSA programming, are discussed in Chapter 7.) Most of these research questions would be appropriate for any component or group of components of JSA programs that ACF would choose to evaluate under any of the delivery modalities or mechanisms.

These research questions follow from the conceptual model in Exhibit 2.1. They consider impacts on participant activities, outputs, and outcomes; they consider both short-term and long-term time frames; and they consider both overall impacts and how impacts vary with participant characteristics and program context.

In considering research questions, it is useful to begin at the far right of the figure, with overall impacts on outcomes. Furthermore, given the conceptual model, short-term earnings is the appropriate focal outcome (although we return to the many other additional, important outcomes in short order).

*RQ0:<sup>4</sup> What is the impact of a JSA program on participant earnings in the months immediately following program enrollment—relative to no government JSA services?*

*RQ1: What is the impact of a particular JSA program component<sup>5</sup> on participant earnings in the months immediately following program enrollment—relative to the current JSA program?<sup>6</sup>*

These two research questions share the same outcomes: earnings in the months immediately following JSA program enrollment, perhaps 6 to 12 months, enough time to participate in JSA services and find—plus hopefully retain for a meaningful interval—the initial job. As discussed in Chapter 4, conventional approaches to the multiple comparison problem (Schochet, 2009) strongly urge choosing a single primary (in the language of multiple comparisons, “confirmatory”) outcome. Multiple comparison considerations suggest choosing between earnings and employment as the outcome of primary interest.

If we would choose a single, most central outcome for the study, it would be earnings, rather than and to the exclusion of, employment or welfare benefits, for example. Of course, earnings gains derive, in part,

---

<sup>4</sup> Question numbering begins at 0 to reflect later considerations that lead to elimination of this first question.

<sup>5</sup> Although we use the term “component” in our research questions, it should be made clear that this can include any dimension of JSA program design/configuration, including the specific components we suggest in Exhibit 2.1, as well as the modes of service delivery within which they cluster. Moreover, various combinations of components, in line with a test of enhanced assistance or enforcement mechanisms, are another possibility.

<sup>6</sup> This question notes, “relative to the current JSA program.” Although the remaining research questions do not explicitly state this, it is implicit that subsequent comparisons are relative to the current JSA program’s configuration, and not relative to a “no-services” or even “business as usual” status quo, where by “business as usual” we mean anything else available in the community.

from impacts on employment, including generating shorter intervals until employment. Earnings, rather than employment, is the most comprehensive measure of labor market success, since it captures not just employment changes, but also changes in hours worked and hourly earnings.

There are other reasons to focus on earnings. First, earnings is the appropriate outcome for a benefit-cost analysis (see Chapter 7). Second, Klerman et al. (2012) suggested that JSA programs might push participants into jobs for which they were not suited, such that a new job would be followed quickly by job loss. Focusing on total earnings over some period of time rather than employment avoids that problem. Third, focusing on earnings, rather than time to employment or other labor market outcomes, gives some attention to job quality (as represented by the hourly wage and hours). Jobs vary in hourly wage and in hours worked. Many jobs for this population are part time. Focusing on time to employment rewards placement in part-time jobs. Fourth, time to employment is not measured for those who do not find a job within the followup window. Econometric methods exist to model such censored data, but using them would considerably complicate the analysis and interpretation of the results. Finally, anticipating some of the discussion of measures and data collection strategies in Chapter 3, earnings are measured, and thus available to an evaluation at low cost from administrative data sources, whereas employment, especially weeks to first job, is not. Measuring short-term time to employment as an outcome would require an expensive survey.

The two research questions differ in what is evaluated. RQ0 asks about JSA compared to no JSA; RQ1 asks about some particular characteristic of a JSA program, implicitly either relative to the current JSA program design or to the JSA program without that particular characteristic. With this study of JSA programs having its primary interest in evaluating the relative effectiveness of particular components and/or modalities of JSA programs, we believe RQ0 would not be helpful. In turn, our designs focus on RQ1, the impact of specific characteristics of existing JSA programs, be it their modes of service delivery or program components.

This primary research question leads to several secondary research questions considering other outcomes within the labor market domain as well as receipt of, participation in, and benefits from various forms of public assistance.

*RQ2: What is the impact of the JSA program component on time to employment?*

One important goal of JSA programs is to get a participant a job as quickly as possible. Time to employment corresponds to that goal. We choose to include this as a secondary outcome for the reasons stated above (and in the footnote): in brief, all other things equal, shorter time to employment will yield higher earnings; but measuring the time to employment is relevant. Together these outcome measures can inform not merely whether a participant gets a job (and how quickly), but also whether that job that will last.

*RQ3: What is the effect of a particular JSA program component on other dimensions of work—in particular, other measures of employment (beyond time to first job), such as employee benefits (e.g.,*

*health insurance, paid vacation/sick time), full-time/part-time, regularity of hours, and regularity of shifts<sup>7</sup>—in the months immediately following program enrollment?*

*RQ4: What is the impact of a particular JSA program component on TANF benefits paid in the months immediately following program enrollment (or being required to participate)?*

*RQ5: What is the impact of the component on other measures of benefit receipt—in particular, SNAP, Medicaid, EITC—in the months immediately following program enrollment?*

RQ1 through RQ5 are deliberately posed in terms of a few months. TANF program rules and TANF program practice imply that JSA programs must be quite short (six weeks or less). Almost all impacts of JSA on time to employment would likely fall within that window. In turn, the primary impacts on other outcomes (e.g., earnings and some broader measures of well-being) will also be in the months immediately following sample entry into a study. However, as discussed in Section 2.2, longer-term impacts are possible. Specifically, the theory of job search, as developed in Klerman et al. (2012), suggests pathways through which impacts might shrink or perhaps grow over time. Consistent with the possibility of longer-term impacts, RQ6 considers a broader time frame.

*RQ6: What is the impact of the intervention on long-term earnings and welfare receipt (and broader measures of employment success and transfer program reliance)?*

Our discussion has focused on impacts on the participant herself (and we noted the possibility of impacts on other family members, including children). In as much as the JSA program has large impacts on employment, general equilibrium effects are possible. Standard economic theory would suggest that pushing more low-skilled workers into the labor market will drive down equilibrium wages for that employment sector. If wages are not flexible (in general, or because of a minimum wage), more effective job seekers may—at least to some extent—displace other workers, with less (perhaps no) increase in total employment and earnings. In turn, an evaluation might explore:

*RQ7: What is the impact of the intervention on nonparticipants through general equilibrium channels (e.g., lower wages due to “flooding” of the supply side of the market)?*

A necessary condition for such general equilibrium effects is a substantial impact on participants. Given that such a substantial impact on participants is an open question, the existence and magnitude of general equilibrium effects are plausibly secondary research questions. Chapter 5 discusses in more detail such potential general equilibrium effects and approaches to detecting them.

The next two research questions do not concern impacts; they are process questions. They consider the inputs to the JSA process on the left side of Exhibit 2.1: engaging participants in program activities, as envisioned by the JSA program designers.

*RQ8: To what extent do participants actually participate in JSA program activities? How does that level of activity compare to the program model and program designers’ expectations?*

---

<sup>7</sup> The likelihood for advancement—in particular, wage and earnings growth—on this job and on future jobs is another dimension of the job. We defer consideration of that dimension to our discussion of long-term impacts in RQ6: What is the impact of the intervention on long-term earnings and welfare receipt (and broader measures of employment success and transfer program reliance)?

*RQ9: To what extent are similar services available outside the program and to what extent do individuals use those services, if not offered them by the program itself?*

RQ0 through RQ9 have been worded as though there are common answers to these research questions for all possible participants. Specifically, answers for the overall population of JSA participants help to address the question whether the programs work, on average, for the target population. However, uniform impacts—on all participants, in all settings—are unlikely. It seems more plausible that impacts will vary along at least three dimensions along the top of Exhibit 2.1 as contextual influences: TANF program characteristics, local economic conditions, and participant characteristics. This introduces the final three research questions for the anticipated study.

*RQ10: To what extent do impacts vary with the parameters of the overall TANF program in the community where JSA services are applied?*

As discussed in Section 2.2, it seems likely that impacts of interventions will vary with the characteristics of the TANF programs in which they are situated. Such arguments suggest that impacts may vary with the details of the broader TANF program. From a design perspective, this suggests testing the intervention in a range of broader TANF program settings.

*RQ11: To what extent do impacts vary with local economic conditions?*

The literature suggests two possible directions with regard to how variation in local economic conditions relates to the impacts of JSA programs. One could argue that, for those individuals who reside in a weak labor market, JSA cannot help because there are no jobs. Conversely, one could argue that in a local area with a strong labor market, participants can find jobs, even without a JSA program. Nevertheless, existing evidence from welfare-to-work programs consistently suggests that the impacts of these programs are pro-cyclical: that more comprehensive programs (though not necessarily JSA) have larger impacts in a strong economy (e.g., Bloom, Hill, and Riccio, 2003; Greenberg et al., 2003). How JSA program impacts vary with variation in unemployment rates should be explored.

*RQ12: To what extent do impacts vary with participant characteristics?*

It is also plausible that impacts vary with participant characteristics; however, the direction of the variation in impact is theoretically ambiguous. One could argue that those with low employability cannot be helped, meaning that impacts will be concentrated on those with strong employability. Conversely, one could argue that those with high employability will find jobs even without help, so that impacts are concentrated among those with low employability. Existing evidence suggests impacts vary with employability (Michalopoulos and Schwartz, 2001; Bloom, Hill, and Riccio, 2003). The existence and magnitude of impact differences by employability—and other participant baseline characteristics—should be explored, with the specific characteristics being identified based on ACF's interest.

Each of these three sources of heterogeneous impacts seems plausible. The first one—variability of impact with TANF program features—seems the most salient. In as much as it is possible to explore the second one, variation with the strength of the economy, doing so would help us to understand the generalizability of study findings in a weak economy, should that prove to be the prevailing climate during the study period, to a stronger economy. The third one, variability with participant characteristics, might be useful to programs as they try to focus resources on those who will benefit the most and tailor services for each participant.

## 2.5 Discussion

This chapter has presented a framework (Exhibit 2.1) for thinking about job search and JSA programs that was developed for the project’s Knowledge Development Report (Klerman et al., 2012, Section 2.1). Using that framework, this chapter considered what to evaluate (Section 2.3) and the corresponding research questions for a future JSA impact evaluation (Section 2.4). Given ACF’s guidance, the evaluation should focus on variations in the configurations of JSA programs that operate through the assistance or enforcement mechanism in TANF programs. Beyond that, ACF is not yet ready to specify which aspects—components and modalities—of JSA programs are of primary interest. Hence, the design effort needs to consider evaluation approaches capable of testing a range of variation in JSA program design and operation.

With respect to research questions, we suggest that the primary research question should concern impacts on earnings of JSA programs’ configuration in the months immediately following program enrollment (or requirement to participate), as RQ1 states. As the remaining research questions pose, the evaluation should consider other outcome measures as well as the influences of economic and policy context and individual characteristics. The boxes at the far right of Exhibit 2.1 suggest that those other outcomes might include TANF benefits paid, broader dimensions of employment and transfer program participation, broader measures of well-being, and long-term impacts. Likewise, the possibility of heterogeneous impacts—by TANF program characteristics, local economic conditions, and individual characteristics—supports inclusion of several additional research questions of interest.

Beyond these impacts on individuals, there is an important research question about general equilibrium effects. Such general equilibrium effects are only likely to be important if substantial impacts on participants exist, so that “ripple effects” into the rest of the local labor market have some salience. We defer until Chapter 5 further discussion of general equilibrium effects and the design implications of attempting to estimate them.

Finally, this discussion has focused on impacts. A study of JSA programs might also include a participation analysis, implementation research, and a benefit-cost analysis. To the extent that these three lines of research can improve the utility and interpretation of findings from an impact study, we discuss relevant design issues in Chapter 7.

### 3. Define Key Measures and Possible Data Sources

Having discussed possible incremental changes to JSA programs that might be evaluated and the associated research questions, in this chapter we consider the key concepts and outcomes and how the project might collect information to measure them. The structure of the chapter follows the project's logic model (see Exhibit 2.1), with successive sections considering: individual context (Section 3.1), outputs and receipt of similar services outside the program (Section 3.2), and short- and long-term outcomes (both in Section 3.3). Then, Section 3.4 considers the issue of multiple comparisons (which we also considered briefly in Section 2.4). The final section provides some concluding comments.

This chapter, and this entire document, proceeds cognizant of the data collection challenges. Anticipating the discussion in the body of this chapter, Exhibit 3.1 provides a high-level summary of data collection options.

Baseline surveys are usually inexpensive per case. In an experimental evaluation, they are usually administered as a prerequisite to service provision and random assignment. There are thus no costs associated with tracking participants; participants are captive within a program in which an evaluation takes place. In contrast, followup surveys are usually very expensive per case. The participant needs to be located, usually after some relatively long period of time. Finding low-income individuals, who are more mobile than the general population, can be very expensive. Inter-survey tracking helps lower those costs. However, even with inter-survey tracking, locating is expensive and incomplete. As a result, high survey response rates (at least 70 percent) are difficult to achieve and make nonresponse bias a plausible concern, especially in the instance where differential response rates between treatment and control groups exist.

Administrative data have a very low per-individual cost, be they from welfare programs or other state or national sources. While administrative data on earnings are of high quality, coverage can be geographically limited (as in multistate metropolitan areas); welfare program administrative data quality varies by state; and structuring and cleaning for the purpose of evaluation analyses can be time consuming.

**Exhibit 3.1: Data Collection Options**

| Type   | Content  | Cost per Case | Notes   |
|--|--|---------------|---|
| Baseline Survey  | As specified by evaluation   | Probably low  | Completion of the baseline survey is usually a condition for enrollment in an experimental study                    |
| Followup Survey(s)   | As specified by evaluation   | High          | Possible response error; in particular, recall error  |
| Welfare Program Administrative Data                        | Some background information, cash benefits, activities (including JSA program) | Low           | Content, completeness, and quality varies by state and data item; requires structuring/ cleaning for evaluation use |
| Administrative Data on Earnings (such as state UI or NDNH) | Earnings   | Low           | Only formal sector earnings and no self-employment earnings.  |

Cognizant of the characteristics of the data collection options presented in Exhibit 3.1, the final section of this chapter builds on the discussion in the individual sections to make some high-level observations about data collection options. In particular, as we discuss elsewhere (in Chapters 4 and 6), (1) JSA

programs are low intensity and are therefore likely to have small impacts; (2) differential impacts across JSA program components/models are likely to be even smaller; and finally, (3) ACF has expressed an interest in understanding differential impacts for subgroups. Each of these steps implies much larger samples than the previous step.

Much of the challenge of designing a data collection strategy is considering how a study might collect data—at feasible cost—for the very large sample sizes that are likely to be needed to estimate plausible impacts. Strategies that are feasible for several thousand individuals may not be cost feasible for ten or a hundred times that many individuals. Based on estimates of specific required sample sizes, we revisit in Chapter 6 options for followup data collection. Information on potential followup survey cost and the importance of the different impact questions the study might address is brought to bear at that point as well. The crucial question becomes: Can the study afford a large followup survey capable of collecting information on any outcome measures of interest? If not, to what extent can the key research questions be addressed without a—or at least without a very large—followup survey?

For now the basic information is sufficient to guide exploration of data collection issues and options in the current chapter.

### 3.1 Individual Context

Assuming an experimental design, information on individual context and background is not strictly necessary in order to estimate program impact. Nevertheless, in an experimental design, individual context—as measured at baseline, before the intervention—has five important uses:

1. **Description of the sample.** For descriptive purposes and to assess external validity, it is useful to be able to characterize the sample's characteristics.
2. **As covariates.** When included in regression models for impact, baseline variables increase precision of impact estimates. In general, lagged values of the dependent variables make the most powerful covariates.
3. **Subgroup analyses.** Individual background characteristics are needed to define subgroups for investigation of differential impacts. Such analyses are useful for addressing the question: What works best *for whom* (see RQ 12)? And the corresponding policy questions: Does any approach dominate any other approach for some observable subgroups (but not for other subgroups)? Do some observable subgroups benefit more from a given approach such that, given limited resources, that approach should be assigned to those subgroups (and not to some other subgroups)? We note that it may not be feasible to implement some such findings (e.g., even if the program works better for some racial or ethnic group, it is probably not legal to make assignments to programs based on race or ethnicity).
4. **Weighting for nonresponse.** No survey has perfect followup. Baseline characteristics can be used to construct weights to correct for differential survey nonresponse (at least in as much as it is correlated with observed characteristics).
5. **Confirming randomization.** A standard check of randomization tests for balance on observed covariates. Although this is not strictly required, it is common practice.

Given those possible uses for individual context, the following list discusses the standard measures of individual context and background. We also briefly discuss sources for questionnaire items (assuming data collection via survey) and briefly discuss data sources.

- **Basic demographics.** These variables include race and ethnicity, gender, marital status, children (number and ages), and education. This information will usually be recorded in the TANF program's administrative data. However, the quality of such administrative data is mixed and often missing. This information can also be easily collected at a baseline interview. Standard questionnaire items exist for these concepts (e.g., the Current Population Survey, the baseline surveys for NEWWS or ISIS).
- **Detailed background information.** Beyond basic demographics, we might also be interested in more detailed measures of background such as intelligence/achievement, measures of personality (e.g., self-efficacy), and measures of preferences (e.g., reservation wage, forward lookingness, relative appeal of cash versus leisure, perceived stigma of welfare receipt). A baseline survey is a conventional way to collect this information. The National Longitudinal Survey of Youth 1997 is a useful starting point for items. The NEWWS baseline survey included information on basic skills, depression, and locus of control. Using NEWWS data, Leninger and Kalil (2008) report statistically significant effects of locus of control on education, but not on employment. They explore, but do not find effects of depression on education, though others have found effects of depression on employment (Michalopoulos and Schwartz, 2001). Wanberg (1997) considers the relationship between job search success and self-esteem, perceived control, and optimism.
- **Motivation.** Vinokur et al. (1995)<sup>8</sup> and Klerman et al. (2012) emphasize the role of motivation to a search for a job. Specifically, Klerman et al. (2012) emphasize that some JSA participants may prefer continued nonemployment to employment, and this may be rational. Pre-TANF evaluations often found that when work increased (so leisure must have decreased, even before taking into account commuting time and time to drop off and pick up children from childcare), total income did not increase substantially (see Klerman et al., 2012, Chapter 2). It is thus plausible that participants vary in the extent to which they are sincerely and actively searching for work (versus minimally and formally complying with program requirements). This distinction is important because oversight of the intensity and, even more, the sincerity of job search is difficult (e.g., does a job seeker "sabotage" job interviews). We have not identified any existing scales to measure the individuals' motivation in the process. Vinokur and Caplan (1987) provide some scales based on the work of Fishbein and Ajzen (1975). Wanberg, Kanfer, and Rotundo (1995) have a slightly more recent exploration of motivation and job search. Wanberg, Zhang, and Diehn (2010) present and discuss a self-administered inventory.
- **Perceptions of Job Search.** In a utility maximizing model, motivation to search would arise from perceptions about the advantages and disadvantages of search, its likely success, and the advantages and disadvantages of employment. Reservation wage (i.e., the minimum wage that a job seeker would accept) is a standard proxy. The NEWWS baseline survey had two questions on

---

<sup>8</sup> See also Caplan et al. (1989), Price, van Ryn, and Vinokkur (1992), and Price and Vinokur (1995).

reservation wage. Wanberg, Zhang, and Diehn (2010) include some sample items on perception and on reservation wages.<sup>9</sup>

- **TANF and other public assistance program history.** Information on JSA participants' past TANF experience—any, timing, duration, and benefits paid—is useful as a covariate, especially in models of TANF receipt—any and amount. Ideally, we would also collect information on other public assistance history, in particular SNAP, Medicaid, and the Child Health Insurance Program (CHIP). For TANF, nearly ideal data are available in the historical files of administrative data systems maintained by (usually state) TANF agencies, though those data vary across states in format, content, and quality. Nearly ideal data on the other programs is usually also available in the respective administrative data systems for those programs. However, those systems are often separate from the TANF program's systems and are often controlled by other state agencies. Thus, gaining access to these data and then processing them is likely to have nontrivial costs. Some information could also be collected via a survey. However, concern about recall bias is likely to limit the detail that can be collected and the quality of what is collected.
- **Earnings history.** Prior labor market experience, including duration of prior employment and earnings history, is available in the historical files of administrative data sources (e.g., Unemployment Insurance or National Directory of New Hires, as discussed later in the chapter). Again, some information could also be collected via a survey. However, again, concern about recall bias is likely to limit the detail that can be collected and the quality of what is collected.

### 3.2 Outputs and Similar Services Outside the System

The intervention is intended to change a JSA program's outputs, i.e., the JSA services received by participants. Thus, an evaluation of JSA program interventions should characterize the JSA services each participant received, including which services were received when, and with what intensity. Exhibit 2.1 lists the modalities of such services and provides some examples. Among the specific JSA services that would ideally be recorded would be: job club, one-on-one counseling, soft skills training and enforcement of job search requirements.

---

<sup>9</sup> We have in mind questions such as:

- “How important is it to the TANF program that you search actively for work?”
- “How likely is it that you could find a job if you looked now?” (very likely, somewhat likely, not too likely)
- “How likely is it that you could find a *good* job if you looked now—where a good job would be full time, regular shift, paying more than \$8.50 per hour?” (very likely, somewhat likely, not too likely)
- “How much better do you think your life would be if you take one of the jobs you could find now?” (much better, a little better, no change, a little worse, much worse)
- “How intensively does the TANF program expect you to search for work?” (response categories to include what is measured in various TANF programs—this one and others)
- “I am concerned about being away from my child” (strongly agree, agree, disagree, strongly disagree)
- “How well can the TANF program monitor your job search?”
- “What is the penalty for failure to comply with the required intensity of job search?” (nothing, they would give me a warning and another chance, they would cut my benefit, they would terminate me from the program for a period of time, they would terminate me from the program forever)
- “If you would not comply, what are the chances that you would be caught?”

This information will usually be recorded somewhere in the TANF program's records. Exactly what information is recorded and the ease with which that information can be extracted for tabulation and analysis appear to vary widely across states<sup>10</sup> and specific data items. Specific issues include:

- Ideally, there would be formatted items in the record (e.g., a record for every one-on-one counseling session—scheduled and conducted; a record for every job club session—scheduled and attended). However, it appears that often this information is only recorded in “caseworker narratives,” such that hand coding would be needed. Such hand coding is likely to be infeasible for large samples of program participants.
- In some welfare-to-work systems, no-shows are a major issue. It follows that the best system would include activities scheduled, whether the individual showed up for and completed the activity and the next step (e.g., reschedule, schedule for an alternative activity, begin sanction process), and the reason for no-show if applicable.
- The extent to which existing codes can be used to identify JSA-specific activities (versus broader welfare-to-work activities) and which type of JSA-specific activities were performed (e.g., counseling versus enforcement).

An alternative would be to ask recipients about services received in a short-term followup survey. Such a short-term followup survey to collect information on services received is common in several related ongoing Abt projects for HHS and DOL. Here we note that with a large sample, such a survey is expensive. However, it seems plausible that samples to estimate differential levels of service received could be smaller than the very large samples needed to estimate impact on labor market outcomes, though tradeoffs exist.

The TANF-based JSA program is often not the only way to receive JSA services. Some studies (e.g., the National JTPA Study) have found that many individuals assigned to the control group receive services similar to those received by the treatment group—but from some source outside the program to which they were randomized (Heckman et al., 2000). In general, no complete (or near complete) administrative data will exist on such services. Thus, a survey would be needed to collect such information. However,

---

<sup>10</sup> Several issues are involved. First is the structure of the computer system used in the operation of the state's welfare-to-work system. That structure determines what information is recorded in the computerized system (versus paper records or never recorded) and how the information is formatted (e.g., a table of date/type of activity/participation versus no-show; or open text “case notes”). For most (but not all) locations, this computer system is state level, which is why we use “state” in the body of the report. However, in some states (e.g., California), substate locations (e.g., counties) use different computer systems. The second issue concerns what is saved. Some systems only keep the most current information, overwriting old information with new information (e.g., the most current appointment). Third is actual practice. A computer system may have a field, but staff may not use it (at all or at least not sufficiently consistently to use the information to track services delivered). Initial scheduled activities may be recorded, but not cancellations and reschedules. Sometimes the reschedules will write over the initial appointment, such that a researcher cannot track no-shows from the computer system. Even if the computer system is common across the state, actual practice may not be. In particular, the completeness of documentation in computer systems is often a function of the nature and quality of supervisor oversight of caseworkers.

such services outside the program seem likely to be less salient in TANF, where the great majority of recipients must be enrolled in some qualifying activity (work, or some program activity).

As to the timing of such a survey, job search and job readiness assistance is limited to the hourly equivalent of no more than 6 weeks per year (12 weeks in states qualified as “needy”). As long as participants move into a defined JSA program promptly after randomization, a followup survey at three months seems appropriate. The earlier the better; the longer the delay until the survey, the more we need to be concerned about recall bias. However, as we discuss in the next section, we might also want to use a short-term survey to collect information on outcomes. Time to first employment is probably best measured in that time frame (i.e., three months after sample entry). However, our recommended primary, confirmatory outcome—earnings—would probably be better measured later (though earnings could be well captured through administrative data).

### 3.3 Outcomes—Short-Term and Long-Term

The evaluation’s ultimate goal is to identify JSA program choices that lead to better client outcomes. We divide those client outcomes into seven groups: (1) attitudinal changes; (2) client job search behavior; (3) labor market outcomes (e.g., time to employment and earnings); (4) program participation; (5) enforcement and sanction experience; (6) broader measures of well-being; and (7) longer-term outcomes (earnings, welfare receipt, and broader measures of family and child well-being).

First, JSA programs might have effects through inducing *attitudinal changes* that lead to more intensive and more successful job search. Such attitudinal changes should begin with intrinsic motivation to search (intensively) for a job. They should also consider extrinsic motivation, i.e., awareness that the TANF program expects job search, the likelihood of success of job search, the positive consequences of work, and the negative consequences of failure to search actively (i.e., sanction). We note that the New Hope study explored, but did not find evidence of attitudinal change (Huston et al., 2001).

Second, JSA programs might have effects through changing *client job search*. Thus, an evaluation could measure client job search behavior. Klerman et al. (2012) suggest a three-step framework for job search: (1) identify job openings, (2) convert job openings into job offers, and (3) decide whether to accept job offers. An evaluation would measure job search behavior at each step of this framework.

The Basic Monthly Current Population Survey (CPS) includes a battery of questions on job search activities that might be appropriate here, including:

- **Active Methods:** contacted employer directly/interview, e-mailed a potential employer, public employment agency, private employment agency, friends or relatives, school/university employment center; sent out resumes/filled out applications, including online; contacted union/professional registers; placed or answered ads, including online.
- **Passive Methods:** looked at ads, including online; attended job training programs/courses.

Given the discussion in Klerman et al. (2012), these questions have three deficiencies. First, they fail to distinguish the three steps of job search. Second, there is no measure of intensity—in this activity or in total time on job search. The National Longitudinal Survey of Youth, 1997 cohort is a promising source for more detailed questions on this topic.

Vinokur and Caplan (1987) provide some scales based on the work of Fishbein and Ajzen (1975). Wanberg, Kanfer, and Rotundo (1995) have a slightly more recent exploration of motivation and job search. Wanberg, Zhang, and Diehn (2010) present and discuss a self-administered inventory.

Third and fourth, we expect the primary impact of JSA programs to be in the short term—*recipients leave welfare faster, work more, and have higher earnings*. Thus, we would ideally measure welfare receipt (any, amount, this spell, all spells), employment (especially time to first job), and earnings (any, amount, this spell, all spells). Here we note several considerations in specifying these outcomes:

- Given that the primary impact is likely to be on short-term outcomes, a short window seems appropriate. The direct impact of JSA programs (i.e., JSA programs provide assistance with job search, leading to employment) will only occur after participation in JSA programs. Earlier, we argued that a followup survey to collect data on program activities should occur perhaps three months after randomization. The primary impacts on time to employment would also be expected to occur during the six-week term of the JSA program, so a survey at three months would be appropriate (assuming prompt movement into JSA after randomization).
- However, in Chapter 2, we argued that the primary outcome should be earnings. Earnings are a flow. We would want to measure that flow over some period of time—a quarter or more. This is especially true since one of the reasons to look at earnings is that they penalize quick job exit. Again, quick job exit would occur after job finding. Thus, in as much as a survey was intended to collect earnings, timing at 6 or even 12 months after randomization would be preferable.
- It appears that some job search programs prioritize any job, such that jobs found have very low wages and very low hours. We care about quick employment because some research shows that it leads to higher earnings. Thus, we probably want to measure earnings. As we argued in Section 2.4, we probably want to prioritize earnings above time to employment.
- The focal population often has short job durations and cycles on and off of welfare, so it seems preferable to use total time in a window, not the current spell (i.e., not time to first job and length of first job, but total weeks employed in a 12-month window).
- Earnings would best be disaggregated into hourly wage and hours.
- Ideally, we would also get a sense of the “quality” of the job, in particular: whether it is a formal sector job (i.e., payroll taxes are paid), employee benefits (e.g., health insurance, paid sick leave and vacation leave), the shift (day versus night), the regularity of the total number of hours, and the regularity of the timing of the hours (e.g., always 9am-5pm versus sometimes 10am-3pm, versus sometimes 7pm-11pm)—with implications for child care and transportation.

Fifth, much of the existing literature can be interpreted as implying that the impact of programs is largely through *enforcement*: participants are sanctioned off of programs, perhaps inappropriately (i.e., when a proper due process system would have established that the participant was actually compliant or had good cause for not being compliant; see the discussion in Klerman et al., 2012). To address this concern, a thorough evaluation would collect detailed information on the sanction process: whether the client exited TANF due to sanction, the type of sanction, the reason for the sanction, and any proxy for the appropriateness of sanction. As of now, it is not clear what might be useful proxies for the appropriateness of the sanction. In addition, we note that what one would want to collect will vary with

the details of sanction policy and process, and what is recorded will vary with the data collection system (as in the earlier discussion about the content of JSA programs received).

Sixth, the broad literature on welfare-to-work programs and the data requirements for a benefit-cost analysis point to the importance of looking at *broader measures of well-being*. Such broader measures of well-being might include:

- Participation in other social programs: SNAP (any and amount), Medicaid, public housing, EITC (any and amount).
- Total household income (i.e., including other transfer payments and earnings of other family members) and income relative to the poverty line.
- Consumption, food security, health insurance coverage, and family and child well-being.

Relatively standard instruments for collecting this information can be found in earlier welfare-to-work surveys and in the Current Population Survey Annual Demographic File (CPS-ADF).<sup>11</sup>

Seventh, in the longer term, the theory of job search suggests that better job search might lead to *better jobs*—higher pay and longer job durations; though Klerman et al. (2012) noted that the evidence for longer job search leading to better jobs is mixed. That report also noted that JSA programs might cut job search short, leading to lower pay and shorter job durations. To test either version of this theory, we would collect data on earnings, welfare participation, program participation, and broader measures of well-being at a longer followup interview. In addition, insofar as JSA leads to better jobs, there might be long-term impacts on broader measures of well-being, including impacts on children.

In summary, we would expect most impacts from JSA programs to occur relatively quickly, and perhaps to fade as those who received weaker JSA also find jobs. Thus, primary interest is probably in outcomes in the first 6 (perhaps 12) months following randomization. Long-term impacts, in both directions, are also possible. Thirty-six months would be a conventional longer followup period.

Ideally, we would collect outcome data at high frequency (i.e., every month or every quarter), and we would collect these data not only for short-term followup (e.g., 6 to 12 months), but also for long-term followup (e.g., 36 months). The challenge is to find cost-effective ways to come close to these ideals, in part by deciding what is crucial and what is less important.

A survey could be used to collect data on any or all of these outcomes. High-frequency survey data collection in conventional telephone mode is prohibitively expensive. Perhaps the study could run a baseline survey, a short-term followup survey (e.g., 6 or 12 months after randomization), and a long-term followup survey (e.g., at 36 months).

Deleting the long-term followup survey or only surveying a subsample of those randomized are options worthy of consideration—options to which we return in Chapter 6 based on more concrete statistical

---

<sup>11</sup> The Current Population Survey (CPS) is conducted monthly. The “Basic” interview is fielded every month and includes questions on demographics and labor market status (including job search activities). In March of every year, the CPS Annual Demographic File includes an augmented set of questions on employment, earnings, and all sources of income in the previous year.

information on required sample sizes. It has to be considered, however, that longer-term impacts are more distal and therefore likely to be smaller. To detect smaller impacts, an evaluation is like to require the full sample.

As we alluded to earlier, some of these outcomes can be measured using administrative data. TANF administrative data are likely to have high-quality and relatively easy-to-process information on TANF benefits paid. Those data are likely to be available on a monthly frequency and with several years of history. From administrative data, we should be able to get long histories of prerandomization TANF benefits (to use as covariates) and high-frequency data to be used as outcomes. However, these data systems vary across states, so an evaluation in multiple states would need to process multiple data systems.

Similarly, there are several possible sources of administrative data on earnings. State UI systems maintain quarterly data on earnings to operate their UI systems. Sometimes, a study can get direct access to those data. If not, ACF's Office of Child Support Enforcement (OCSE) aggregates state UI data, augments those data with earnings data from federal employment, and organizes the combined quarterly data on earnings into a consistent format, lowering processing costs, through the National Directory of New Hires (NDNH). Access to the data is limited (some variables are suppressed or recoded), but that access should be sufficient for this study. The U.S. Census Bureau's Longitudinal Employer-Household Database (LEHD) also aggregates state UI data. That is an alternative data source. Finally, the Social Security Administration collects separate, but annual, data on earnings in order to operate the Social Security and Medicare programs. Again, access is limited, but these are a third data source. In summary, good administrative data are available on earnings, with some history (two years in the OCSE data) and high frequency (i.e., quarterly) going forward. Note, however, that there is no information on other dimensions of employment—hours, hourly wage, benefits, working conditions, even precise timing of employment (within the quarter).

Beyond these TANF benefits and earnings data, obtaining administrative data on outcomes seems even less feasible. It may be possible to get administrative data on SNAP benefits. Getting access to those data is likely to require considerable effort (and is not assured), and state file formats will differ. As the number of states in the evaluation grows, the cost/benefit analysis of acquiring such data looks less attractive. We have not identified promising sources of administrative data on other outcomes.

### **3.4 Multiple Comparisons and Confirmatory Outcomes**

Modern discussions about statistical testing note that we rarely test a single outcome. Instead, evaluations perform many tests. It is, therefore, useful to think about the tests—and, in particular, the probability of Type I error, i.e., concluding that there is an impact where there is no impact—as a group. This is known as the “multiple comparisons problem” (e.g., Schochet, 2009; see also the discussion below in Chapter 4 of this document).

The standard solution to the multiple comparisons proceeds in three steps:

- Prespecify a set of “confirmatory” outcomes.
- Report statistical tests viewing those confirmatory outcomes as a group.
- Report other statistical tests, but deem them “exploratory,” where exploratory outcomes receive less weight in the writeup of the results.

This standard solution appears to be an appropriate approach for this evaluation. As we argued in Section 2.4, given the evaluation’s logic model and the likely impacts of JSA programs and their components, earnings in the 12 months after randomization appears to be an attractive (single) confirmatory outcome.

### 3.5 Discussion

The previous sections have considered the various measures implied by the logic model and touched on issues of data collection. Exhibit 3.2 attempts to summarize the data collection discussion. For each major measure, we consider possible measurements through: A baseline survey, a short-term followup survey (perhaps at 6 or 12 months), a long-term followup survey (perhaps at 36 months), and administrative data.

**Exhibit 3.2: Data Collection Strategies**

| Concept                     | Data Collection Strategy |            |           |                   |
|-----------------------------|--------------------------|------------|-----------|-------------------|
|                             | Baseline Survey          | Short-Term | Long-Term | Administrative    |
| Baseline demographics       | ++                       | +          | +         | +                 |
| Preprogram earnings         | ++                       |            |           | ++                |
| Program activities          |                          |            |           | ++ (program data) |
| Postprogram employment      |                          | ++         | ++        | ++                |
| Welfare participation       |                          | +          | +         | +++               |
| Other program participation |                          | ++         | ++        | ++                |
| Income relative to poverty  |                          | ++         | ++        |                   |

NOTE: “+” could provide this information; “++” is a good source for this information; “+++” is a very good source for this information.

A long-term followup survey is conventionally scheduled at 36 months (e.g., NEWWS). Thirty-six months is long enough for the participant to have concluded job search and found a job, and to see if that job lasted and whether there was earnings growth. Longer followup might be even better, but we do not want to further lengthen the evaluation timeline.

The earlier discussion noted two possible timings for a short-term followup study. If the short-term followup study is focused on program activities (and participation in similar activities outside the program), then 6 months appears to be an attractive followup point. However, if the short-term followup survey is intended to measure outcomes (e.g., employment, welfare participation, income), then 12 months appears to be an attractive followup point. If the survey is intended to collect information on both program activities and outcomes, then it will need to occur at the later time (e.g., 12 months). Fielding the survey at 12 months rather than 6 months will probably have some negative effect on the quality of data on program activities.

Exhibit 3.2 is intended to emphasize the stark tradeoffs in data collection. Per case, survey data are expensive; administrative data are not. With a caveat about the varying formats and quality of state TANF data, the very highest priority outcomes—earnings, benefit receipt, and perhaps sanctions—are easily measured using administrative data from the welfare program itself and from UI records.

However, other important outcomes are not recorded in administrative data. In particular, it does not appear to be possible to construct a broad measure of household well-being from administrative data. Some sources of income may be available in easily accessible administrative data (e.g., SNAP payments,

receipt of Medicaid, perhaps UI benefits), but others probably not (e.g., EITC payments, earnings of other household members such as partners or cohabiting family members, child outcomes). For those who leave TANF, the household structure recorded in the TANF administrative records may no longer be correct, so it is not possible to compute a poverty measure. Broader measures of well-being such as consumption and food security are not available in these administrative data sources.

Thus, the evaluation design has three broad options, all of which assume administrative data on earnings and TANF participation:

1. ***Large samples and survey data collection.*** This option leads to very large—perhaps infeasible—evaluation budgets.
2. ***Smaller samples and survey data collection.*** This option leads to imprecise estimates, and probably the inability to detect plausible impacts of a JSA program on the confirmatory outcome—earnings. This option would almost certainly lead to an inability to detect differential impacts of JSA programs.
3. ***Large samples but less than universal survey effort.*** This option would rely primarily on administrative data, with the size of the survey adjusted to yield a feasible budget. However, the survey sample—and thus the sample for many important, if secondary outcomes—will be smaller and power, therefore, lower, which might not necessarily be a problem for some measures (such as service receipt) but would be for others (such as food security).

Which option to choose will depend on the available budget and the relative importance placed on measuring various outcomes. Our sense is that an evaluation is only worth doing if it has power to detect (at the very least) impacts of the magnitude that JSA might generate or that would be of policy relevance. The power analysis in Chapter 6 suggests that doing so will require very large samples. As discussed there, detecting these impacts may push us towards Option 3, above, with a relatively small sampling fraction, relying predominantly on administrative data to estimate impacts on the confirmatory outcome of interest.

## 4. Design Options for Measuring Causal Impacts on Individuals

The primary goal of this document and contract is to provide design options for an evaluation of JSA programs that will answer the research questions posed in Chapter 2. As such, this chapter is the core chapter. It discusses specific evaluation designs, and related considerations, that may be appropriate for evaluating the effectiveness of a wide range of job search programs and activities for specific target populations.

The evaluation, by necessity, will involve multiple study locations. Several factors necessitate this: variation in JSA program configurations, both as they exist and as they might exist in an evaluation context; ability to answer research questions about the role of variation in local conditions and policy context; and sample size demands, be they individual-, subgroup-, or site-driven.

This chapter is organized as follows. First, we discuss why and how a study of JSA might consider external validity: while the policy evaluation field generally prefers designs that ensure internal validity, we urge considering external validity as well, including not just generalization to wider policy-relevant populations and settings, but also to desired policy decision-making times. We then connect those considerations to issues of multisite evaluation research. After briefly introducing the basics of the randomized experimental evaluation design, we turn to individual- and cluster-level randomized designs, presenting the contrasts that they permit evaluating. Next, we discuss the general analytic framework for these designs and suggest some additional analytic issues that are relevant.

In general, this is a theoretical chapter, describing design options independent of their feasibility. That said, we conclude this chapter with a brief summary identifying the key challenges to planning an evaluation that would be based on any of these designs. We then revisit and elaborate on these challenges within a framework for assessing the tradeoffs across the many factors in an evaluation of this sort.

### 4.1 External Validity

This section discusses how the settings, target populations, and timing of a JSA impact evaluation will affect generalization of the study findings to a larger, more relevant policy universe—i.e., it addresses the evaluation’s *external validity*. After an examination of the goal and challenges of generalizing from impact evaluations of social programs, we consider design and analysis approaches for a JSA evaluation that could increase the external validity of the study’s findings. Similar considerations apply whether dealing with individual-level or cluster-randomized designs.

#### 4.1.1 The Goal and the Challenge

The concept of external validity in policy evaluations arises not from an abstract notion of some “universe” of potential study domains that could have been included in the research. It arises from the very practical issue of how policy makers will use a given evaluation’s findings. To identify the right external validity *goal* or *target*, one begins by asking: Which group of job seekers, in which communities, and in what years will be affected by the policy choices the study findings are expected to influence? If, for example, ACF anticipates federal policymakers deciding whether to launch a nationwide JSA program for unemployed low-income parents in 2018 based on the results of an evaluation that emerges from this design work, one would like to develop a design (impossible though it may be) that gathers data from a representative sample of 2018’s or 2019’s unemployed low-income parents across all geographic corners of the country. This would constitute ideal external validity, were it doable. The information from such a

study would extend results into the larger “external” environment where the policy choice is situated: the information would concern the very group of citizens (or, really, a statistically representative sample of the very group of citizens) being judged as the nation makes its 2018 policy choice.

Alternatively, this design work could be envisioned as supporting an evaluation that informs state government policy choices concerning the allocation of workforce development funds that state legislatures and governors control to JSA strategies to certain populations of importance to individual states. While this would again mean, in the (impossible) ideal, getting evaluation data on future members of those populations for the time period in which state choices based on the evaluation would be activated, it would have quite different implications for the geographic and demographic coverage needed to attain external validity for the alternative domain of policy choices the research is to guide.

The policy context of the intervention also affects external validity. For national policy purposes, ACF will benefit most if the JSA interventions studied are embedded in a range of state TANF program environments. The conceptual framework in Chapter 2 also points to the importance of the local economy in conditioning the way JSA impacts participant outcomes. This aspect of context also needs broad representation that is characteristic of the nation overall to make the findings of the evaluation as valuable as possible to national policy.

This discussion highlights two of what we see as four major challenges in attempting at least some degree of external validity in the DOSE impact evaluation design should ACF set external validity as one of its goals:

- The right external domain requires anticipating the evaluation’s use in future policymaking, and multiple domains could be relevant.
- Research on the effectiveness of a current policy intervention can never directly address the domain of policy interest, which concerns the future.

The other two major challenges are:

- Covering the broad range of settings and target populations of interest to an impact evaluation in a representative, statistically adequate way requires very large samples of locations and individuals, and therefore large funding for the study.
- One must achieve high levels of cooperation among the selected locations for the study (sites) once a statistically representative sample of communities is identified for the evaluation.

We discuss our approach to addressing these challenges in the JSA evaluation context.

### **4.1.2 Design Responses: Choosing or Approximating a Representative Sample**

In the face of these daunting challenges, what can be done in the design phase—and what analytic techniques can be planned—to produce more externally valid findings for the policy choices a JSA impact evaluation is likely to guide? Major work on this question is underway or has recently emerged from the evaluation methods literature, including a set of papers by Olsen and colleagues (Bell et al., 2011; Olsen et al., 2012) and others (e.g., Peck and Mayo, 2011; Tipton and Hedges, 2011).

Olsen et al. (2012) emphasize statistical probability sampling of sites, and note that this goal has proven attainable in several larger-scale randomized impact evaluations over the last 15 years.

Probability sampling from the population of all TANF programs in the nation is likely unrealistic for a future JSA evaluation,<sup>12</sup> and in any case, does not make findings statistically generalizable in a formal sense to the *future* population of citizens about whom one most wants to know the tested intervention's effectiveness. Failing this, the best practice is to broaden the range of contexts covered on the dimensions the conceptual framework suggests are likely to matter: state TANF program parameters, job seekers' characteristics, and local labor market conditions. If possible, one would like to stratify on key variables in these domains and fill each cell with at least a portion of the research sample. As long as no cells are empty, the analysis can make adjustments to simulate a different, perhaps more generalizable, mix of contextual factors, in order to provide more representative impact estimates. We acknowledge that, indeed, the number of cells to consider would be large, suggesting that some narrowing and prioritizing should be involved early in the design phase to select which factors are most essential here. We recommend these elements be included in carrying out any future JSA impact evaluation involving TANF recipients and potential TANF eligibles.

### 4.1.3 Analysis Responses: Enhancing External Validity with a Nonrepresentative Sample

As suggested in the previous subsection, analytic methods are available to translate impact findings from nonrepresentative sites into measures of greater salience for the nation's TANF system as a whole—if not in a formal statistically representative way, at least to increase the “face external validity” of the results.<sup>13</sup> Many researchers are currently working to determine the best performing of these methods, where a truly externally valid impact estimate is available as a benchmark for judging method performance; findings of this research could inform the choice of the best method for contextualizing study findings by the time a future JSA impact evaluation reaches the analysis stage.

## 4.2 Multisite Considerations

With discussion of external validity as motivation, we now turn to a brief discussion of how ACF might envision assembling a collection of sites in which to test the effectiveness of various JSA program features. The rationales for including many sites in this evaluation include the following:

- **Program variation.** Depending on the specific JSA intervention(s) that ACF chooses to evaluate, it seems unlikely that a single given site, or even a small number of sites, would be able to provide or create the program-specific contrasts of interest.
- **Contextual variation.** To answer questions about the role of variation in policy context and economic conditions, there would need to be variation along these dimensions. A single site, or a small number of sites, would not be able to provide suitable variation.

---

<sup>12</sup> Special types of TANF JSA program environments are needed to test one particular JSA strategy against another in a multisite study. As discussed elsewhere in this document, only locations with at least one of the desired JSA models in place can be considered, leaving out of consideration large parts of the nation's TANF population. Neither this report nor its partner (Klerman et al., 2012) document the national distribution and configuration of these JSA programs, and doing so would need to be an early step in any evaluation effort.

<sup>13</sup> Among the methodologies available are regression with treatment interaction terms (from which predictions for populations with different characteristics can be calculated), reweighting cases stratified by baseline characteristics to match the population of interest, and use of propensity score weights.

- **Sample size.** In order to detect the differential effects of variation in JSA program elements—be they service provision modalities, components, or combined features that characterize the underlying mechanisms through which JSA operates—a single site, or a small number of sites, is unlikely to be able to provide sufficient numbers of observations (as elaborated in Chapter 6).

We will revisit these issues in the concluding chapter, but mention them here to emphasize the need for an evaluation to create one or more *contrasts* that will be its focus. Here we suggest a range of possibilities that selected (systematically or randomly) sites would have to undertake to create a contrast that would support these experimental designs:

- **Near nothing.** Identified sites would need to be willing and able to ration access into multiple existing program modalities/components (that are sufficiently distinctive from one another to create a useful contrast). They would not necessarily have to change anything about the configuration of JSA offerings within their locations except the process by which individual job seekers gain access to particular parts of the program.
- **Scale back.** Identified sites would need to be willing and able to reduce availability to some existing program modality/component (to create a useful contrast) and ration access into it and the rest of the program.
- **Scale up.** Identified sites would need to be willing and able to add/increase some existing program modality/component (to create a useful contrast) and ration access into it and the rest of the program.
- **Scale back and up.** Identified sites would need to be willing and able to reduce availability of some existing program modality/component while also adding/increasing availability of some other existing program modality/component (to create some useful contrasts) and ration access into each of these arms as well as into the rest of the program as it currently operates. This strategy has the potential to create what we refer to later as a “maximum variation” study.

We will revisit these, and other, site-related issues later, specifically with reference to cost and administrative practicality. For now, we simply establish the general idea that multiple sites will be needed, and sites will need to be willing and able to create an evaluable contrast. We acknowledge that to “create an evaluable contrast” that involves scaling back, as we have just described, may involve reducing the availability and types of services for some types of individuals who would previously have had access to them. This demands consideration of the ethics of doing so. Consistent with general principles of research ethics, any evaluation design should allocate services in a fair way, and avoid withholding services of known effectiveness solely for the purpose of evaluation. Because these services are not an entitlement, there is no legal restriction in reducing what some individuals can access. Further, because no directly applicable evidence exists on the effectiveness of these specific JSA services, it is not clear that reducing what some individuals can access is a real disadvantage. That said, some TANF program rules require individual participation in job search, and so individuals must not be disadvantaged by having fewer services available to them.

We do not intend to suggest that to create an evaluable contrast sites should or would do anything much different from their current configuration of JSA, unless they would be especially interested in doing so. That is, while we imagine that some locations might be especially interested in innovating (making them different from non-innovators), it is our expectation that—given how engrained JSA is in TANF programs—sites are more likely comfortable with their JSA programs. As such, any modifications that

we would recommend as part of an evaluation would be based on the existing configuration of programs, with changes coming predominantly from changing the process by which individuals (a) gain access to particular program components or (b) experience particular program components—including the overall intensity of the contrasted JSA interventions. By making modifications to programs’ flow and access structures, an evaluation would be able to create a contrast or series of contrasts that supports evaluating the relative effectiveness of particular program elements. We return to this point after introducing the basics of experimental design.

### 4.3 Internal Validity

Strategies for measuring policy impacts should first focus on assuring unbiased estimation of the average effect of the intervention on the intervention participants examined. This property—that the statistical expectation of the impact estimate equals the true average impact of the intervention on the study sample—is called internal validity. Its assurance in the current study is the focus of this section.

The concepts of cause and effect are the basis of inquiry in the field of policy or program evaluation. As policymakers and program administrators establish, implement and evolve public policies and programs, determining the extent to which those policies or programs *cause* changes in outcomes is essential to justifying ongoing funding or suggesting policy changes (or termination). Because policy interventions are taking place in an environment of change in a variety of contexts, the primary way to isolate the *effects of the policy* or program is to create a control group that represents what would happen in the absence of the policy but under the same environmental changes (what evaluators call the “counterfactual”). Randomization is the key mechanism for creating a reliable counterfactual, one that allows researchers to net out the influences of any other plausible, rival explanations for observed differences between with-program and without-program outcomes in the focal population. The strength of an experimental evaluation design—which relies on the random assignment of subjects to treatment and control status—lies in allowing only the intervention to vary systematically between treatment and control groups. The difference in average outcomes between the two groups nets out all other non-chance differences that could be mistaken for program impacts. Being able to interpret the difference as the effect of the intervention is important for program administrators and policymakers alike; it is certainly preferable to arguing over whether a difference in outcomes has other origins. Being confident that impacts accrue *because of* a program or policy focuses energies on the right information. Although some nonexperimental designs can provide a solid basis for causal inference under certain circumstances, we argue against exploring nonexperimental design options. Instead, we start from the position that the best option for this study is an experimentally designed evaluation where internal validity is unassailable, and discuss various options for using this design.

***What “status quo” does the control group represent?*** In general, in order to know whether a particular treatment has an impact, we compare it to the counterfactual, what would have happened in the absence of the treatment. This type of control group represents the status quo or “business as usual.” In field experiments such as this, the control group rarely gets “no services” but instead consists of control group members accessing whatever services are otherwise available in their community. The contrast that this creates involves answering: What difference, relative to the existing constellation of services, does the program we’re evaluating make? As well as being feasible to create, this contrast is usually the most policy-relevant: government support of job search is important, but not the only means for people to secure support. We would not expect private and nonprofit efforts to disappear, and so government efforts must be evaluated against the backdrop of existing services. That said, these existing services might not

reach the specific targets, which provides justification for government intervention in the first place. This is the contrast that is represented in RQ0 in Chapter 2.

Another variant of the status quo involves not what else is out there in the community, but rather, the immediate, existing configuration of services from which we might learn about the relative effectiveness of selected components or against which an alternative, or specific modality/component, might be tested. This approach—letting the counterfactual represent an existing program—is especially useful for understanding what impacts variation in JSA program features might have. In the JSA context, this might mean taking away or adding some program component in order to detect the extent to which its presence/intensity really matters to improving outcomes for job seekers. As noted in Chapter 2, we do not focus on a status quo that represents no JSA, or even whatever other JSA is available in the community.

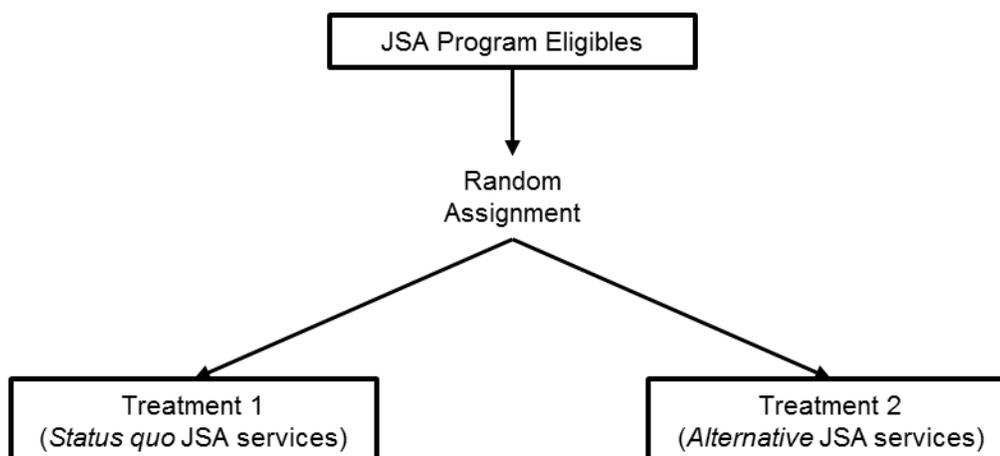
From here forward, when we refer to the “status quo” we are referring either (a) to the existing constellation of JSA services that a given TANF program offers or mandates, or (b) some developed configuration of JSA services that creates a contrast to an alternate set of JSA services. For example, if the evaluation team identified 20 sites that share roughly the same package of JSA, and if sites were amenable, program eligibles could be randomized to access the package as is, *less* one component, with the remainder of program eligibles being randomized to access the package *including* the component of interest. This created contrast would allow estimating the effect of that specific component of JSA within the context of an existing program. It will be the practical challenge of a future evaluation to identify the appropriate sites, programs, and JSA components that can be configured to create useful, evaluable, and policy-relevant contrasts.

### 4.4 Individual-Level Randomized Designs

This section describes experimental design possibilities in which *individual job seekers* are the unit of randomization. Policy experiments commonly randomize individuals to one treatment group and a control group. In this case, we explicitly state that the “control” group is really a second treatment arm. One can think about randomizing individuals to multiple treatment arms as well. Doing so permits estimating the relative effects of various policy/program elements. Multiple treatment arms are attractive for the evaluation of programs with multiple components where a primary question of interest is the effect of specific program components. Specifically, this section describes three variants of individual-level randomized experimental designs: the basic (two-arm) randomized design, a multi-arm randomized design, and a factorial design. For each, the discussion offers one or more illustrative examples of how the method might be used specifically in the job search context. Other JSA-specific applications are possible.

#### 4.4.1 Basic Randomized Experimental Design

**Design.** In its most basic form, a randomized controlled trial randomly assigns individuals—program eligibles—to treatment or control status. Those in the treatment group gain access to services, those in the control group do not. In this case, the control group represents the standard configuration of JSA services, or a modified configuration of JSA services that creates a contrast of policy interest between the treatment and control groups. Exhibit 4.1 provides a graphical depiction of the process for allocating a research sample between treatment and control status. Among JSA program eligibles, random assignment channels some into gaining access to the existing configuration of JSA program services (the “control” group), and others into gaining access to an alternative configuration of JSA program services.

**Exhibit 4.1: Example of Basic Individual Randomized Design**

**Example.** This basic design would be appropriate for testing the effectiveness of any given alternative JSA intervention relative to the existing program configuration, as indicated in Exhibit 4.1. An alternative JSA treatment could add, take away, or change specific components of the JSA program in order to create a contrast that allows us to detect the *difference* that the addition/removal/change makes, relative to the existing program. In order to support analysis of this latter contrast, sites would need to be selected to change their program in some predetermined way, such as adding soft skills training in settings where such training does not currently exist, an approach sometimes described as an intervention “bump up” strategy. Then, individuals would be randomized to have access to the standard JSA program as it currently exists (this would be the control group/status quo), or to have access to the JSA program *plus* soft skills training. The resulting treatment-control contrast provides an estimate of the impact of adding soft skills training to the existing configuration of services where such training was not already in place. One might also imagine removing some JSA program component in order to create a contrast between two versions of JSA services, where the difference between the two groups’ outcomes estimates the impact of the specific component removed (present for one group but absent for the other). Several other possible applications of this type of evaluation design exist to test the effectiveness of specific JSA program components.

To estimate the effects of a singular programmatic change relative to the existing JSA program’s design, reasonable consistency would need to exist across multiple sites in both the change and—to a lesser degree—the base program for optimal learning.<sup>14</sup> This design would be most appropriate if ACF knew which single program component were of greatest interest to evaluate.

#### 4.4.2 Multi-Arm Randomized Experimental Design

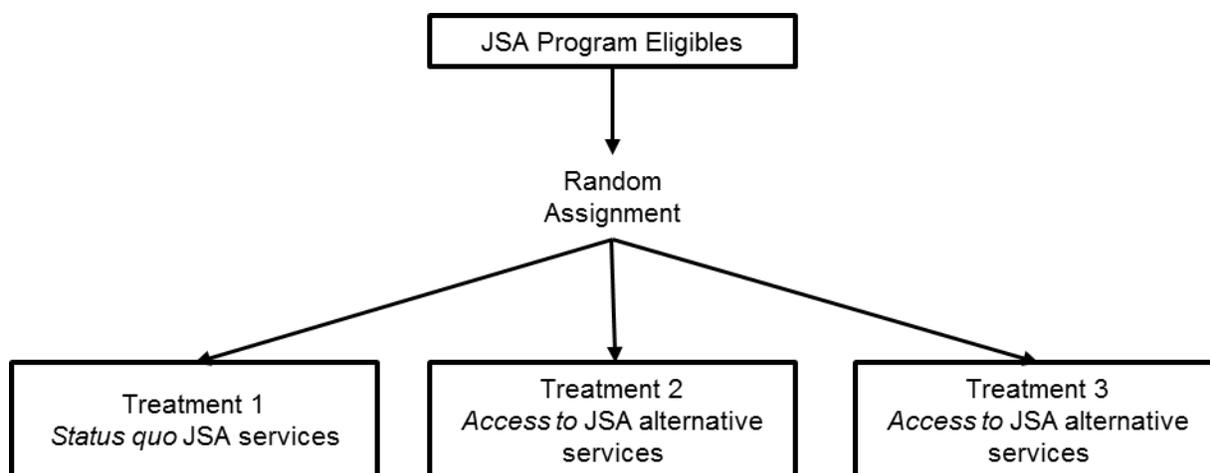
The two-arm randomized design can test a single JSA component relative to the existing configuration of services. In contrast, a multi-arm design allows parsing effects of variation in program components by experimental means. As such, it is preferable to the two-arm design for understanding impacts of more

<sup>14</sup> Diversity in the base program, so long as the change is uniform, could strengthen the study by making its findings regarding the impact of the change generalizable to more contexts.

than one individual program component. The next subsections discuss two variants of a multi-arm experiment: the first is a three-arm trial, and the second involves a factorial design.

**Design.** A standard three-arm design randomizes to three distinct treatments, creating a contrast between two selected components and the existing JSA program. As before, what might be thought of as the control group receives the existing configuration of JSA program services and the treatment groups would involve one of two alternative versions of or components of JSA. As shown in Exhibit 4.2, the treatment groups represented by “Treatment 2” and “Treatment 3” (specific variations of JSA programs) are compared to “Treatment 1” (the existing JSA program), and they can also be compared to each other to provide an estimate of the differential impact of the two program variants, however defined. As we elaborate below, a viable example might be to compare a basic JSA program (without access to job club or use of individualized assessment) to a JSA program where enrollees participate in job club as part of their intervention or where enrollees’ program experience involves individualized assessment. These treatment-treatment differences can inform the impact of including one of these specific program features as part of the constellation of JSA services provided. This design can involve more than two alternative treatment arms as well, to the extent that there are particular program components that can be distinctly provided, without contaminating the experience of individuals in the other treatment arm(s).

**Exhibit 4.2: Example of Multi-Arm Individual Randomized Design**



**Example.** For example, if ACF were interested in testing the effectiveness of having access to job club and the effectiveness of having access to individualized assessment, then the three groups in a multi-arm trial would be (1) standard JSA program without job club or individualized assessment, (2) standard JSA program with job club, and (3) standard JSA program with individualized assessment. The particular locations selected to operate this sort of trial could be those in which neither component currently exists but where they are willing and able to add these program components in conjunction with rationing access to them, at least for the period of the evaluation. In contrast, selected test sites could be locations where both alternative program components exist and sites are willing to ration access to each in order to test the relative effectiveness of each program component; under this scenario, some individuals would be deferred from accessing services to which their predecessors had access, at least for some period of time. In brief, having multiple treatment arms must be feasible from a program management perspective: program staff must be able to provide variants of their program to the individuals assigned to receive each variant (preferably with different staff implementing the different variants, if staff can be assigned to

variants at random), without risking crossover of individuals across treatment arms. We believe that, in the JSA context, this could be straightforward and therefore a relatively low-risk proposition.

#### 4.4.3 Randomized Factorial Design

Given ACF’s interest in exploring the relative effectiveness of the various modalities of JSA service provision and of various JSA program components, a more complex design, known as a randomized factorial design, might be more powerful. In a factorial design, selected treatment options or “factors,” each of which has two (or more) levels, are combined, where the levels can be low or high dosage or simply absence or presence of the factor. As Collins et al. (2005, pp. 65–66) elaborate, “which program components are working well; which should be discarded, revised, or replaced; which dosages of program components are most appropriate; whether delivery components are enhancing, maintaining, or diluting intervention efficacy; and whether individual and group characteristics interact with program or delivery components” are important questions—usually well more important than the average treatment effect—that a factorial design can aid in answering. With these questions central to ACF’s desire to understand *what* it is about JSA programs that is working, or not, or could be refined to be more effective and efficient, this design approach seems especially appropriate.

**Design.** In its simplest form, the factorial design varies two treatment dimensions or factors, randomizing to each individually and to both together. If the levels of each factor include “absence” or “presence,” then the absence of both factors represents a *status quo* control group, as signified in Exhibit 4.3. The design in Exhibit 4.3 can accurately be considered a four-arm experiment, where two alternative JSA service models (“Treatment 2” and “Treatment 3”) offer access to one of two distinct alternative intervention options and the third alternative JSA service model (“Treatment 4”) offers access to both of the alternative options.

**Exhibit 4.3: Factorial Design with Control Group**

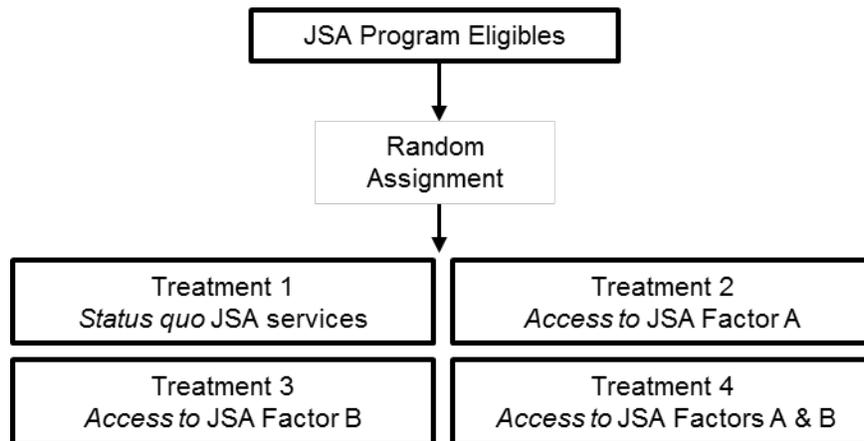
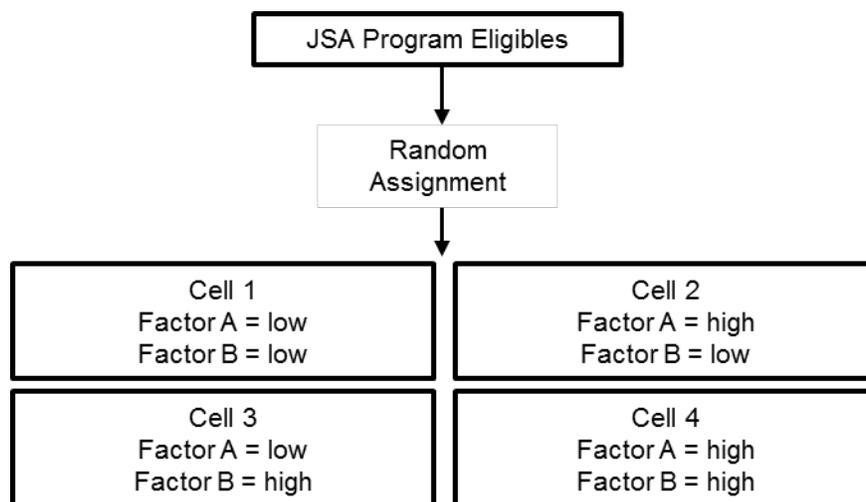


Exhibit 4.4 presents a more traditional factorial design, where levels of intensity are varied at random.

**Exhibit 4.4: Factorial Design with Varying Levels of Intensity**



This design offers greater power to detect the impact of each factor’s overall impact because it has a relatively larger “control” group for a given factor, which includes all of those randomized not to have access to that factor. Referring to Exhibit 4.4, for example, to estimate the effect of a higher intensity of Factor A, we compare the combination of cell 2 and cell 4 against the combination of cell 1 and cell 3. To estimate the overall effect of a higher intensity of Factor B, we compare the combination of cell 3 and cell 4 against the combination of cell 1 and cell 2.

As Shadish, Cook and Campbell describe a factorial design (2002, p. 264), “each participant does double duty” by the nature of his/her exposure to particular factors or factor intensity levels in the multi-cell designs in Exhibits 4.3 and 4.4. In addition, these designs are configured to support estimation of the incremental (marginal) impacts of individual factors or their intensities and to estimate the combined impact of the two factors or their intensities—an analysis that can reveal whether the combined factors are synergistic or, together, less than the sum of their parts. These additional questions do not come at a sample size premium, as shown in Chapter 6, for the main reason just mentioned.

A 2x2 factorial design involves four cells but can support answering eight questions. Among them, for example, the design represented in Exhibit 4.4 can inform the extent to which Factor A at high intensity, Factor B at high intensity, or the combination of Factors A and B at high intensity is more effective than having both factors at low intensity. A 3x3 design involves nine cells and supports answering 16 questions. A “fractional” factorial design might actually test only a subset of the possible hypotheses available from the full matrix, though having empty cells reduces power. A well-known application of this design is the New Jersey Negative Income Tax experiment in the late 1960s and early 1970s, where income guarantees were defined by poverty level (at 50, 75, 100, and 125 percent) and the “tax rate” on earnings (30, 50, and 70 percent). Although this 4x3 design would have 12 possible cells (and the ability to answer many more questions because of the various interaction effects), researchers assigned participants to the eight least expensive and most politically palatable combinations.

**Example.** To follow our earlier example, one could imagine randomizing job club and individualized assessment as two factors that a 2x2 factorial design could test, both individually and in combination with

one another. As before, the sites that would be appropriate to take on such a test would need (a) to have neither option in place and be willing to introduce them, or (b) to have these options in place and be willing to scale back on them, both while randomizing access to them, individually and together. Sufficient administrative and programmatic capacity is needed to carry out this design, but certainly the benefit of learning about the relative effects of program components both overall and incrementally promises to inform ACF about how best to design JSA programs along these dimensions. Given ACF's interest in the variation in service provision modalities and selected JSA program components, a many-celled factorial matrix could be designed, considering specific cells (a fractional design) to optimize what ACF might learn.

Experimental designs that capitalize on having multiple treatment arms—with individual job seekers independently randomized among the treatment options—will be the best way to capture the relative effects of specific elements of a multifaceted treatment. As the discussion so far has noted, the selection of sites to support an evaluation of JSA service modalities and program components will be an important choice. While it would be ideal to randomize essential program features in order to estimate their effects without bias, substantial management capacity would be necessary to carry out a many-celled factorial design, for example. It would seem ideal to recruit capable sites to do so. But, another possibility is to consider recruiting specific sets of sites that, for reasons of their own, already vary program features for different participants. For example, if ACF were interested in understanding the relative effectiveness of job club within program settings where programs emphasize group approaches, as opposed to the effectiveness of job club within program settings where programs emphasize individually directed approaches, then an evaluation might select sites that specifically differ along these dimensions of service provision and then randomize all individuals within those settings to have access to job club as part of their treatment or not. That is, all individuals within sites that emphasize group approaches might be randomized to have access to job club or not independently of how they are currently assigned; and all individuals within sites that emphasize individual approaches might be randomized to have access to job club or not. This variation in program design (emphasis on group vs. individual approach) would be the criterion for selecting sites to participate in the study. To the extent that a sufficient number of sites meet the criteria of interest, sites themselves could be randomized to take on a specific program modification, a topic to which we turn next, in discussing cluster-randomized designs.

### 4.5 Cluster-Randomized Designs

Cluster- (or group-) randomized designs are those that use, as the unit of analysis, an aggregate of individuals, such as a classroom or school, or a program site or state. It is the aggregate unit itself that is randomized, and all of the individuals accessing treatment in that group receive the same treatment, with the control group being comprised of a site that is randomized out of treatment including all of the individuals therein.

Cluster-randomized assignment raises feasibility issues. In order for a cluster-randomized design to be feasible, the aggregate units—we will call them “sites” in this context—need to be willing to let some outside entity dictate the nature of their service provision or program design, i.e., whatever the result of the “coin toss.” Usually, administrators or managers at sites themselves are the ones to decide the configuration of and rules associated with services offered; but site-level randomization requires that sites be willing and able to implement their program according to an external dictate. In some naturally occurring clusters, this requirement is not a major burden: for example, schools that are evolving their curriculum might dictate that certain classrooms or teachers use a particular textbook or approach. In

social services more broadly, state-level policy may dictate some program rules, but local-level programs may decide more specifically how they implement those rules, to the extent that local discretion is allowed or feasible. We have a cross-state natural experiment in effect in the U.S., where our federalist system results in substantial state-level variation; but that variation is not random. To incorporate random variation into the evaluation of social policy, where site-level randomization occurs, requires that states choose to impose on substate units the randomization to treatment or control status of various program features or components that are of sufficient policy relevance as to be evaluated via an experimental design.

The main analytic reason to adopt cluster design is a concern about interference. What happens in one arm of the randomized design may affect what happens in the other arm if they take place in the same community and labor market—at least to some extent. There is concern about spillovers of interventions across randomization arms, such as in the example of job development below. In this and other cases with important spillover potential, individuals within each site must all be assigned to the same treatment status.

Another reason for preferring a cluster-randomized design is the relative ease of randomizing, compared to individual-level randomization. With individual randomization, study implementers must design a random assignment process that fits the special circumstances of each site, train intake staff of the program agency on random assignment procedures, and continuously monitor randomization over a many-month period. There are also many more ways that execution of random assignment results—who gets which of the randomized interventions—can be subverted. Moreover, with individual-level randomization, program staff need to build a new step for identifying who from among their eligibles “wins” the lottery into their existing processes. It is not necessarily difficult to do so (and we have well-established tools for doing it), but as the number of sites increases, the number of process variations that need to be made more comparable in order to accommodate randomization increases. In turn, site-customized work to embed random assignment demands evaluation resources but is needed to obtain comparable contrasts across randomization arms in various sites. Group randomization does not require within-site changes to accommodate individual-level randomization; instead, sites need only focus on the treatment modification that is part of the evaluative test. Moreover, sites may not have the scale or the management sophistication to run two different programs at once.

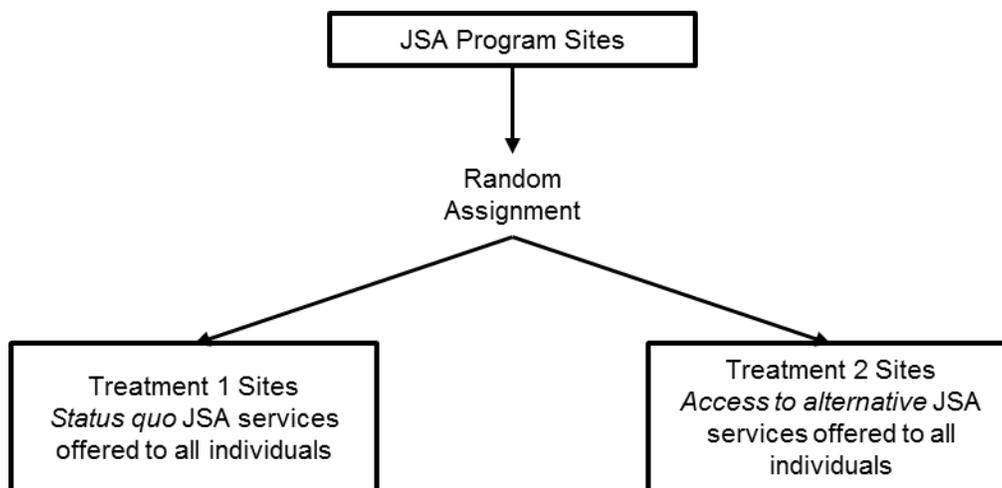
In individual-level designs, job seekers could be randomized to alternative interventions. But not all interventions are randomizable at the individual level. Moreover, implementing two (or more) interventions in a single site—and keeping them separate—can be difficult from a management perspective. For example, job development, including engaging with employers for various reasons, cannot be expected narrowly to affect only some treated individuals. If employers become better able and more interested in hiring the kinds of individuals coming from these job search programs, then they are also more likely to hire control group counterparts, thus attenuating measured impacts of the JSA on worker outcomes. The opposite is not true: any program alternative that could be randomized at the individual level remains a candidate for cluster-level randomization.

### **4.5.1 Randomized Sites (Without Individual Within-Site Random Assignment)**

The cluster-randomized design without individual within-site random assignment is optimal when, as noted above, there is concern about spillovers of interventions across randomization arms.

**Design.** A standard cluster-randomized design randomizes entire sites to each treatment condition as shown in Exhibit 4.5. The unit of analysis remains the individual job seeker, however. As before, individuals in the treatment group would receive access to augmented or expanded program services, which would involve adopting some additional program component(s) beyond the basic package of services. The control group would be one where individuals receive the existing configuration of JSA program services available in their site. The difference between this and individual random assignment is simply that in a cluster-randomized design, all individuals in a site are assigned to the same treatment condition.

**Exhibit 4.5: Basic Group-Randomized Experimental Design**

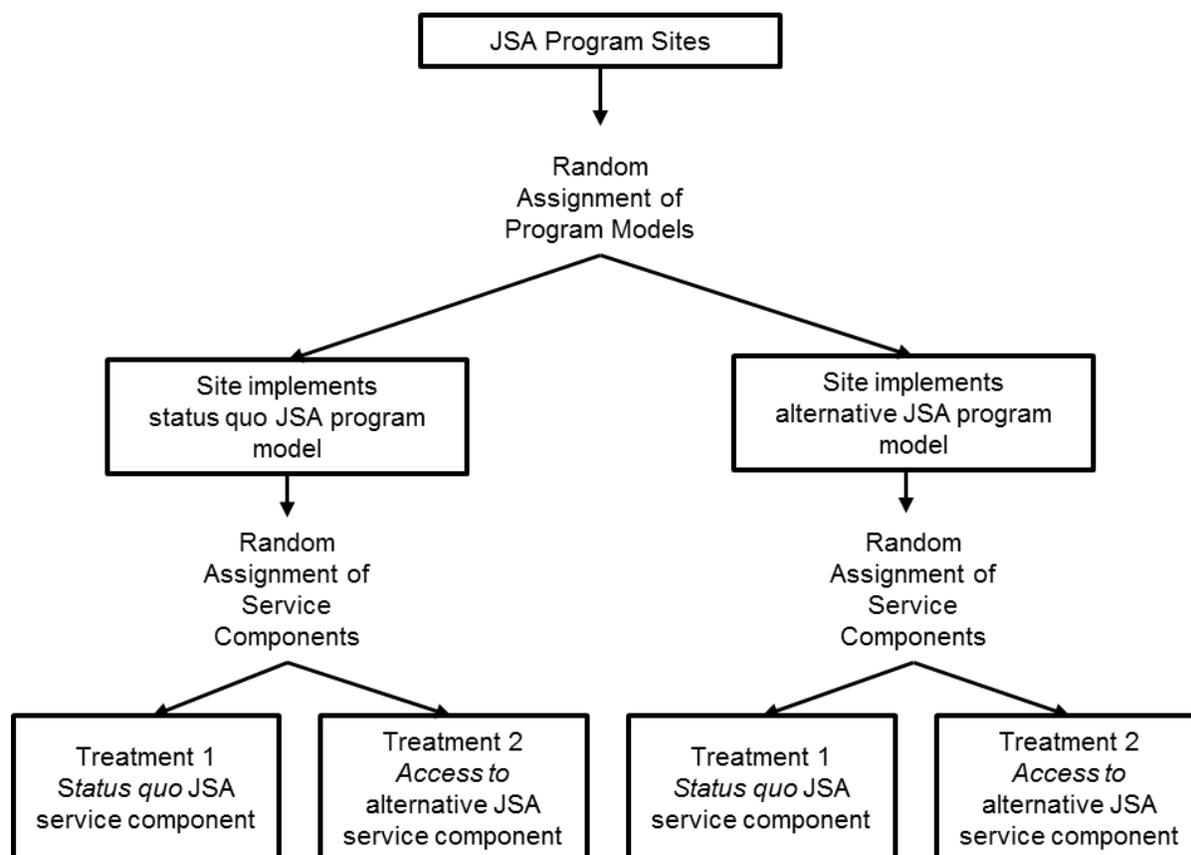


#### 4.5.2 Randomized Sites Plus Individual Random Assignment to Components within Sites

Some intervention models are best randomized for whole sites rather than individuals within sites, such as job/employer development. Important “spillovers” to all TANF job seekers can arise when working through community channels of this sort, contaminating the policy comparison when two sets of individuals in the same sites are randomized and compared. For site-wide JSA strategies, it is better to rely on random site-to-site variation than naturally occurring variation, since the latter can correlate with other factors that affect JSA impacts and confound the pure effects of JSA program differences. It is also important, as discussed in Chapter 6, to consider site-level randomization for the purpose of measuring the general equilibrium effects of a given TANF JSA effort on the local labor market as a whole.

Following randomization at the site level, further randomization of individual TANF recipients within sites yields further policy information on alternative JSA strategies that are appropriate for individual-level variation, such as overall intensity of the intervention or use of particular intervention strategies such as individual-level counseling, thereby leading to interest in “double randomization.”

**Design.** In the double-randomized design, sites are first randomly assigned to one of two (or more) JSA program models; then, within each site, individuals are randomly assigned to specific program dimensions or components, forming status quo and alternative JSA service groups. This structure, depicted in Exhibit 4.6, allows for unbiased estimation of program impacts at both levels, to inform two different policy questions: how are outcomes different by program model at the higher level and how are they affected by status quo versus alternative service components at the lower level?

**Exhibit 4.6: Double-Randomized Experimental Design**

**Example.** This evaluation design would be most appropriate for a simultaneous test of JSA program features that have spillover effects, such as employer/job development, *and* of JSA program components that can be randomized effectively at the individual level. The former would be randomized at the top level of the flow in Exhibit 4.6, where a given site implements a particular JSA program model. Then, at the next level, individual JSA participants are randomized into or out of specific services such as one-on-one counseling or intensive job club. In general, the design simultaneously provides information on more policy options with a given set of sites and individuals—but with less statistical precision (due to smaller sample sizes) for each policy comparison. Moreover, if the *same* service components are randomly assigned among individuals in multiple sites, then this design can reveal how impacts of those features vary depending on the site-level JSA model in which they are embedded, free from confounding by pre-existing differences across sites.

#### 4.5.3 Methods for Estimating Impacts

This section lays out the analyses that estimate impacts across the individual- and cluster-level designs that we have discussed.

**Analysis for the Basic (Two-Arm) Randomized Experimental Design.** The difference between the mean outcomes of the alternative treatment group and the mean outcomes of the status quo treatment group is an unbiased estimate of the alternative treatment’s impact. Conventionally, impacts are estimated within a multiple regression context in order to increase the precision of the estimates. Baseline variables are used

as covariates to increase the precision of the impact estimate, thereby reducing the size of the effect that a given sample size and design can detect.

The regression model is specified as follows:

$$Y_i = \alpha + \delta T_{2i} + \beta X_i + e_i \tag{Eq.1}$$

where:

$Y$  is the outcome of interest;

$T_2$  is a dummy variable that indicates assignment to the group receiving access to the alternative JSA services in Exhibit 4.1 ( $T_2 = 1$ ) rather than the status quo ( $T_2 = 0$ );

$X$  is a vector of individual background characteristics (both here and in subsequent equations);

$e$  is a random error term (both here and in subsequent equations); and

the subscript  $i$  indexes individuals (both here and in subsequent equations).

In this model, the intercept,  $\alpha$ , represents the regression-adjusted mean outcome for the status quo treatment group. The  $\beta$  coefficients noted above are not of substantive interest. The coefficient  $\delta$  is interpreted as the impact of being in the alternative treatment group, which is known as the ITT (“intent to treat”) estimator, on which we elaborate, including alternatives, later in this chapter.

**Analysis for a Three-Arm Randomized Experimental Design.** The impacts that this design supports estimating are the following: T2-T1, T3-T1 and T3-T2. The impact estimating equation, modified slightly from that in Eq.1, adds an indicator for the second treatment arm, as follows:

$$Y_i = \alpha + \delta_1 T_{2i} + \delta_2 T_{3i} + \beta X_i + e_i \tag{Eq.2}$$

where:

$Y$  is the outcome of interest;

$T_2$  is a dummy variable that indicates assignment to the group receiving access to one type of alternative JSA services ( $T_2 = 1$ ) in Exhibit 4.2, rather than status quo JSA services or the other type of alternative JSA services ( $T_2 = 0$ ); and

$T_3$  is a dummy variable that indicates assignment to the group receiving access to the other type of alternative JSA services ( $T_3 = 1$ ) in Exhibit 4.2, rather than status quo JSA services or the other type of alternative JSA services ( $T_3 = 0$ ).

As above in Eq.1, in the model presented in Eq.2, the intercept,  $\alpha$ , represents the regression-adjusted mean outcome for the status quo treatment group. The  $\beta$  coefficients noted above are not of substantive interest. The coefficient  $\delta_1$  is interpreted as the impact of being in treatment arm 2—in this example having access to job club—and  $\delta_2$  is interpreted as the impact of being in treatment arm 3—in this example having access to individualized assessment. One can test the hypothesis that  $\delta_1 = \delta_2$ .

**Analysis for Randomized Factorial Design.** As noted in our discussion of the randomized factorial design, eight questions can be addressed by this design. In terms of the factors depicted in Exhibits 4.3 and 4.4, they are:

- What is the overall impact of Factor A (or Factor A at high intensity versus low intensity)?
- What is the impact of Factor A alone?
- What is the overall impact of Factor B?
- What is the impact of Factor B alone?
- What is the impact of Factors A and B together?
- What is the difference in the impact between Factor A alone and Factor B alone?
- What is the additional impact of Factor A added to Factor B?
- What is the additional impact of Factor B added to Factor A?

The power of the factorial design comes from its use of those randomized not to receive Factor A as the control group against which one can compare the outcomes of those assigned to Factor A, for example. This is what is represented in the “overall” impact estimate. This is in contrast to the impact of Factor A “alone,” which uses only the control group’s mean outcomes as the counterfactual and thereby has the same power as a similarly sized two-group design.

The general regression model associated with a simple 2x2 factorial design is as follows:

$$y_i = \alpha + \delta_A F_{Ai} + \delta_B F_{Bi} + \delta_{AB} F_{Ai} F_{Bi} + e_i \quad (\text{Eq.3})$$

where

$y$  is the outcome;

$F_A$  is a dummy variable for the presence of Factor A (or of high-intensity Factor A) in the assigned intervention (= 1 when Factor A is present, = 0 otherwise); and

$F_B$  is a dummy variable for the presence of Factor B (or of high-intensity Factor B) in the assigned intervention (= 1 when Factor B is present, = 0 otherwise).

As with the previously specified models (in Eq.1 and Eq.2),  $\alpha$ , the intercept, remains interpreted as the mean status quo control group outcome. One can add baseline covariates, which serves to increase the precision of the impact estimates, as in a conventional impact estimating model; and, again, interpreting their coefficients is generally not of interest. Most centrally, the  $\delta$ s represent the impact of their corresponding factors:  $\delta_A$  is the mean impact of factor A;  $\delta_B$  is the mean impact of factor B; and  $\delta_{AB}$  is the added synergistic effect of combining Factors A and B in addition to the simple additive contribution.

***Analysis for Site-Level Randomized Design.*** When sites are randomized, impacts can still be estimated as the difference in mean outcomes between individuals in the two sets of sites created through randomization, similar to an individual random assignment design. However, because the unit of analysis for the outcomes of primary interest (e.g., employment and wages) is at the individual level, but program characteristics are randomly assigned and adopted at the site level (and thus will affect all participants within these units similarly), the standard individual-level comparison will incorrectly estimate the standard error of the impact estimate. This problem is widely recognized in the literature (Moulton, 1986, 1990; Raudenbush, 1997; Raudenbush and Bryk, 2002).

A common approach to dealing with the effect of clustering individuals within sites when randomizing whole “clusters” (i.e., sites) is known as multilevel modeling (sometimes called HLM, hierarchical linear modeling). When the time comes for impact analysis, two equations are specified, one for individuals and one for sites. The following is a stylized version of a multilevel model that can be used to estimate impacts in a cluster-randomized design with correct standard errors. In this approach, individual outcomes are modeled at Level 1 while Level 2 models site parameters as the unit of random assignment.

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \sum_{k=1}^K \beta_{kj} X_{kij} + \varepsilon_{ij} \quad (\text{Eq.4})$$

$$\text{Level 2: } \beta_{0j} = \gamma_0 + \gamma_1 T_j + \sum_{m=1}^M \gamma_m + \mu_j$$

where at Level 1 (the individual level):

$Y_{ij}$  is the outcome of interest (e.g., employment) for individual  $i$  in site  $j$ ;

$X_{kij}$  is the  $k^{\text{th}}$  baseline characteristic for individual  $i$  in site  $j$  (e.g., =1 for males);

$\beta_{0j}$  is the regression-adjusted (for the  $X$ s) mean value of the outcome measure in site  $j$ ; and

$\varepsilon_{ij}$  is the residual error for individual  $i$  from site  $j$ .

At Level 2 (the level of randomization):

$T_j$  is a dummy variable equal to 1 if site  $j$  is assigned to the alternative JSA services model and 0 if assigned to continue with the *status quo* model;

$\gamma_0$  is the global regression-adjusted mean of the outcome measure;

$\gamma_1$  is the coefficient of interest, which represents the estimated impact of the various JSA models compared to one another across all sites; and

$\mu_j$  is the residual error for site  $j$ .

A positive and statistically significant  $\gamma_1$  coefficient shows the alternative JSA intervention to be more effective than the status quo JSA program on average across sites (90 percent of the time, anyway, should one use an alpha of 0.10). Note that cluster-randomized designs are not limited to two treatment arms. Sites could be the unit of random assignment in a multi-arm studies as well.

An alternative to using the multilevel model structure and assumptions is to estimate a conventional OLS model with cluster-robust (heteroskedasticity-consistent) standard errors. Since information about the variance components does not seem to be of interest, this approach is simpler and therefore likely preferable.

**Analysis for the Double-Randomized Design.** Multilevel modeling is also a suitable analytic approach for analyzing data from a double-randomized design, where the impact of the JSA service contrast randomized to individuals within sites (e.g., intensive job club) is first estimated for each site, and then the impacts are compared across sites assigned to each of two site-level intervention models (e.g., with and without employer/job development). A simplified version of the multilevel formulation for estimating impacts in the double-randomized design is given by the following, which are estimated simultaneously to obtain correct standard errors:

$$(1) Y_{ijt} = \alpha_j + \delta_{0jt} Q_{ijt} + \sum_{k=1}^K \theta_{kj} X_{kij} + \varepsilon_{ij} \quad (\text{Eq.5})$$

$$(2) \alpha_j = \alpha + e_j$$

$$(3) \delta_{0j} = \gamma + \beta_0 T_j + v_j$$

The variable  $Q_{ij}$  is a dummy variable indicating the treatment from within-site randomization for individual  $i$  in site  $j$ . The coefficient  $\delta_{0j}$  thus represents the comparative impact of JSA services randomized in site  $j$ . The variable  $T_j$  indicates the site's JSA intervention model (=0 for status quo, and =1 for alternative). Also, in the second equation  $\alpha_0$  is the global regression-adjusted mean outcome. The coefficient of interest at the site level,  $\beta_0$ , represents the differential impact of program model A relative to program model B. Covariates may be added to the model to increase precision, as in the cluster-randomized design (represented by  $X$ ).

As noted above with respect to Eq.4, this analysis may just as effectively be executed using conventional OLS regression with cluster-robust standard errors. Given that the variance components might not necessarily be of keen policy interest, the more straightforward approach may be preferable.

#### 4.6 Additional Analytic Methods

Above, we have described the basic analysis one undertakes in estimating program impacts for each design. Fortunately, the simplicity and strength of the experimental design means that nothing especially sophisticated is needed. That said, some additional analytic considerations warrant discussion. This subsection discusses five specific additional analytic considerations, which refer to (1) no-shows and crossovers; (2) exogenous subgroup analyses; (3) individual variation in program experience; (4) cross-site variation in program design; and (5) multiple comparisons considerations.

**No-Shows and Crossovers.** As defined earlier, the ITT represents the average difference in outcomes between those in the treatment group and those in the control group, regardless of whether they actually participated in the program being offered. As such, it may dilute the effect of receiving program services because treatment group outcomes average those who took up the offer, those who engaged completely in the program, and those who did not participate at all.

While the ITT estimate provides the impact of being assigned to the treatment group (regardless of whether one actually participate in the program), the “treatment on the treated” (TOT) estimate provides the impact of taking up the offer to participate in the program. TOT estimates are computed by dividing the ITT estimate by the proportion of treatment group members who actually participate. As noted by Bloom (1984), if no-shows can be assumed to experience none of the impacts, then the correction attributes the difference between outcomes for treatment and control group members to the fraction of treatment group members who actually received treatment. In this case, we do not believe the assumption of no effects on no-shows holds and so do not recommend pursuing a TOT analysis that applies the Bloom estimator. Because of the mandatory nature of job search in the current TANF context, those who do not engage in JSA will certainly be affected, at the very least by experiencing program sanctions. That is, with the main assumption of computing the TOT not being met, the TOT estimator may be especially irrelevant in this case. If ACF were interested in pursuing an analysis of the differential effect of no-shows in this context, however, then we might suggest an alternative analysis—such as that established in Peck (2003)—of post-random assignment participation choices of experiences.

In addition, the existence of crossovers in the control group—those who find their way into one treatment despite having been assigned to the other—will bias differential impact estimates downward by making

the two experimental arms more similar in terms of average JSA experiences. True crossovers in a two-arm experiment, where the control group represents no services or business as usual, are generally avoidable, but we point to Angrist, Imbens and Rubin (1996) and Angrist (2006) as established approaches for taking crossovers into consideration in analysis. This work would be appropriate for consideration in a design where the two groups each receive different variants of JSA services and crossovers might be less easy to avoid than in the case of having a nonprogram control group. In terms of multiple treatment arms, a literature is only now emerging for addressing this difficulty analytically.<sup>15</sup> These methods, once more fully developed, certainly should be brought into the analysis phase of any randomized impact evaluation that grows out of the current design effort.

**Exogenous Subgroup Analysis.** In response to Chapter 2’s RQ12, we discuss how to analyze program impacts by discretely defined exogenous subgroups, those identified by some baseline (prerandomization) characteristic. A split sample subgroup analysis does just what the name suggests: splits the sample into its subgroups of interest (for example, men or women; those with or without a high school degree at baseline) and analyzes impacts within each group. This “unconditional” subgroup analysis uses the same regression model as states for the basic individual-level impact estimating equation:

$$Y_i = \alpha + \delta T_i + \beta X_i + e_i \tag{Eq.6}$$

where each subgroup is analyzed separately, and where:

*Y* is the outcome of interest; and

*T* is a dummy variable that indicates treatment group (=1 for treatment; =0 for status quo).

The intercept,  $\alpha$ , represents the regression-adjusted mean outcome for the status quo treatment group, and the coefficient  $\delta$  is interpreted as the impact of being in the alternative treatment group for the subset of the sample analyzed. An alternative to the approach presented in Eq.6 is to conduct a “pooled” or “conditional” subgroup analysis, where an interaction term in the model captures the effect of being in the selected subgroup and experiencing the treatment, as follows:

$$Y_i = \alpha + \delta T_i + \gamma T_i S_i + \beta X_i + e_i \tag{Eq.7}$$

where:

*Y* is the outcome of interest;

*T* is a dummy variable that indicates treatment group (=1 for treatment; =0 for status quo); and

*S* is a binary subgroup indicator (1=for in subgroup; 0=for not in subgroup).

In this case,  $\alpha$  and  $\delta$  are interpreted as above: the status quo treatment group mean and average program impact, respectively. The coefficient in the term that interacts treatment with subgroup,  $\gamma$ , is interpreted as the incremental effect of being in the subgroup and experiencing the alternative treatment.

**Multiple Comparisons Considerations.** A final important analytic topic to raise is that of the multiplicity of hypothesis tests that come along with a multi-arm trial such as this, especially when there are multiple

---

<sup>15</sup> For example, the Family Options Study being conducted by the U.S. Department of Housing and Urban Development, which randomized homeless families among three different interventions (Abt Associates, 2012).

outcomes and/or multiple subgroups of interest. That is, conventional hypothesis tests assume each test is independent; but in this situation, where we would be evaluating three contrasts, so that a 5 percent probability becomes a 14 percent probability of incorrectly concluding an impact has occurred when none has. That is, the probability of Type I error—the chance of wrongly concluding that an impact or difference in impact exists in one or more of the comparisons, when in fact one does not—increases as more hypotheses are tested. If interested in two outcomes across three tests, the chance of making a Type I error increases to 26.5 percent. Further, if we were interested in knowing these impacts across two subgroups (for two outcomes across the three T-C comparisons), the chance of making a Type I error for one or more of the findings rises to 46.0 percent.

Given the current fluidity of the difficult and much-debated issue of limiting the potential for false positive results in impact evaluations and how this might best be done statistically, what will constitute the best approach will depend on other specific features of what ACF decides to evaluate and with what design. We advocate choosing a single confirmatory outcome, which we earlier argued should be earnings and which later justifies our decisions in estimating minimum detectable effect sizes accordingly. For other circumstances (e.g., exploratory tests and subgroup analysis), this report does not explore the varied options for adjusting for multiple hypothesis tests, though many exist (e.g., Bonferroni adjustment, Benjamini-Hochberg, step-down methods). Instead, we recommend that ACF address the multiple comparison issue directly after it makes other key design decisions, recognizing that issues of multiplicity will have implications for those decisions, primarily as relevant to sample size needed to support detecting effects of a given magnitude.

### 4.7 Conclusion

This chapter has discussed individual- and group-level randomized designs that support estimating the causal effects of variation in JSA program configurations, be they the modalities of service delivery or specific components of an intervention. It presented the analytic approach to estimating program impacts under each of these designs as well as information on some additional analytic considerations. It also discussed some ways in which an evaluation intended to estimate the impacts of JSA programs might be designed to increase its generalizability and policy relevance. Importantly, we believe that feasibility—including getting sites to agree to adopt and to implement a configuration of their existing services that creates an evaluable contrast—is a major challenge, which we revisit in this report’s conclusion.

## 5. Design Options for Measuring General Equilibrium Effects

When policymakers consider one model of JSA programs versus another model of JSA programs, they usually start by considering impacts on participants in the programs. But, JSA programs may also have impacts on other populations, such as employers, other workers who may be displaced from jobs by JSA participants, and those affected by any market-level consequences of JSA such as changes in wage rates and product and service prices (Calmfors, 1994). Those impacts should also be considered in constructing an overall assessment of the desirability of one JSA program model versus another.

This chapter considers whether such “spillover” or “displacement” effects—a subclassification of “general equilibrium effects”<sup>16</sup>—will be fully captured by the data collection and analytic strategies discussed to this point and, if not, how the design can be modified to measure and estimate such impacts. The chapter starts with a discussion of the displacement problem as it has been framed in the labor economics literature. We then consider possible approaches to estimating displacement effects, first by adding to the individual-level random assignment impact design discussed in the previous chapter and then by looking inside that randomized design for evidence on spillover effects.

### 5.1 General Equilibrium Effects in a Job Search Context

Typical random assignment evaluations of job search programs estimate the average impact on participants by including some or all members of the eligible population in the research and assigning some included cases to treatment and others to control. The impact of the program compared to no JSA is then estimated as the mean difference in outcomes between treatment and control group units. The conventional individual-level random assignment evaluation does the assigning at random and uses the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1974, 1977), which asserts that outcomes for any one (treatment or control) sample group member do not vary by their treatment (or control) assignment. For example, the outcomes for a given treatment group member receiving JSA, such as getting a job (or not), have no effect on (and are not in turn affected by) anyone else in the research sample. If SUTVA holds, then the estimated impact can be thought of as the improvement in earnings or employment for participants, relative to what their outcomes would have been in the absence of the program. This paradigm can be extended to the comparison between two (or more) levels or types of JSA rather than JSA versus no JSA.

In a typical design, individuals outside the random assignment sample are not considered. This can be problematic. To see why, consider the case outlined by Meyer (1995, p. 107) in which the local labor market has a fixed number of jobs. Because of this limit, improvements in job search skills for some workers that lead to greater employment for those individuals necessarily leave fewer jobs for those outside the program. So other individuals who would have worked suffer a job loss. In this example, conventional random assignment methods (or, for that matter, nonexperimental methods) overstate the net employment and earnings impacts of the more successful version of JSA: they measure the positive

<sup>16</sup> General equilibrium effects related to scale-up of program coverage to a larger population, and the resulting labor market job seeker/quantity changes and their social welfare consequences, are not a focus of interest in the current evaluation (since TANF JSA programs will already be operating at their natural full scale in the evaluation sites should the study be implemented).

impacts on that intervention group, but the negative impacts on the rest of the labor force—including the experiment’s untreated control group—are not measured.

The example of a fixed total number of jobs is extreme, but it seems likely that displacement of non-JSA workers will occur. For that reason, the typical partial equilibrium estimate of impact (i.e., the one provided by the random assignment sample with individual-level random assignment) taken on its own does not represent the most meaningful impact estimate.

Scholars who have previously scrutinized this issue argue that potential general equilibrium effects shift how we should approach impact analyses of labor market interventions. In particular, several studies have found that displacement of untreated job seekers can account for a substantial portion of the gains to JSA participants. For example, in a study of a reemployment bonus program, Davidson and Woodbury (1993) find that displacement of UI-ineligible workers constitutes 30 to 60 percent of the gross employment effect of the bonus program. More dramatically, Lise, Seitz, and Smith (2004) find that general equilibrium effects actually reverse the benefit-cost conclusions implied by a partial equilibrium experimental evaluation of the Canadian Self-Sufficiency Program, but these authors consider how overall labor market outcomes might change if the intervention studied were taken to a much larger scale—not the perspective to be adopted by the current study, in which only a limited number of TANF recipients can be brought into the studied JSA interventions when operated at the scale actually used in the random assignment evaluation. In a more recent example, Crépon et al. (2012) find that the positive impact of a JSA program on job seekers relative to untreated job seekers came partly at the expense of other workers; the program reduced the relative job-search success of nontreated job seekers. This was especially true in weak labor markets with substantial competition for jobs. The authors conclude that “the main effect of the program was to help those treated to find a job slightly faster, at the expense of others who subsequently took longer to find employment.” The examples outlined here show that not only is the displacement effect large enough to dramatically alter conclusions regarding the effectiveness of a job search program, but also that the effect is measurable for actual-scale implementations of JSA efforts.

In light of the potential for displacement or spillover to nonresearch sample workers, it is essential to develop estimation strategies that deal with situations in which workers who are not in the randomly assigned treatment group for the evaluation are affected by the intervention. We consider several such methods here, some from the literature and some original.

## 5.2 Possible Approaches to Estimating General Equilibrium Effects

There are several research strategies that have been developed to estimate the general equilibrium effect of labor market interventions. Which one(s) to adopt will depend on the goals of the JSA study, the particular objectives of the progenitors of these methods, and the resources available to the research team. Possible approaches include conducting larger-scale studies (Banerjee and Duflo, 2008), such as clustered random assignment designs; combining experiments with quasi-experiments (Gautier et al. 2012); double randomization (Kremer and Muralidharan, forthcoming); and structural modeling (Lise, Seitz, and Smith, 2004). This subsection explains each of these approaches in detail and weighs their pros and cons in relation to the anticipated JSA impact evaluation.

### 5.2.1 Cluster-Randomized Designs

In situations where there are spillovers (e.g., displacement effects), the conventional approach is to randomize entire groups of individuals to the two experimental groups, where the two groups encompass the individuals among whom spillovers may occur. This approach is becoming common in the evaluation of educational interventions, where teachers, schools, or even districts may be the unit of random assignment although the unit of analysis is the student (see, for example, Raudenbush, 1997). The parallel in a JSA context is that the unit of random assignment encompasses the whole group of competing job seekers for a given set of jobs, i.e., the entire low-skill local labor market (LLM).

If the goal of the study is to estimate the overall impact of a program on both individual- and market-level outcomes, compared with a counterfactual in which a different program is implemented, this estimation strategy is the “cleanest” of any discussed here for estimating equilibrium effects related to worker displacement marketwide, and is the easiest to explain to a nontechnical audience. The design is simple random assignment, but with assignment occurring at the level of the LLM instead of the individual level; i.e., a sufficiently large number of LLMs would be selected for the evaluation, and a portion would be randomly assigned to implement each of the two JSA program models to be evaluated.<sup>17</sup> Outcomes would then be measured for a representative sample of job seekers in each LLM (regardless of whether those job seekers actually *used* program services); this encompasses the population of all job seekers in the LLM. Hence, any displacement effects are accounted for in the difference in mean outcomes of job seekers in the different LLMs and the correct net impact is estimated.

As a simple example of how this design accounts for displacement effects, consider Meyer’s (1995) scenario outlined in section 5.1. In this scenario, the number of jobs in the economy is fixed, and the program simply shifts jobs from some job seekers to others—i.e., the net impact on the number of job holders and earnings in the economy is zero, but this is overstated in an individual-level random assignment design where the control group consists of job seekers who were negatively affected by the program. In a cluster-randomized assignment design, on the other hand, the control group consists of job seekers who are *not* affected by the program, because they are in different LLMs. When employment in LLMs assigned to implement the program (which include equal numbers of people helped and people hurt by the program) is compared with employment in LLMs assigned to the control, the estimated net impact will be zero, which is equivalent to the true net impact on employment in this scenario.

When the research goal is to estimate the difference in impacts between two program models, the LLMs each would be assigned to one of the two program models—with no need for a no-services control. Likewise, the outcome of interest could include any outcomes measured at the individual level (e.g., earnings). The clustered design also has the advantage that the impact on market-level outcomes can easily be estimated when those measures are available, which is not straightforward (or even possible) using individual-level random assignment. For example, we might theorize that by better matching job seekers with suitable jobs, a new program may increase efficiency in the local economy—which could be measured as a change in prices for certain goods or services (or perhaps an increase in the total number of advertised job openings, or some other measure). Using a cluster-randomized assignment design, this

---

<sup>17</sup> Note that one of the program models could be a no-services control if the research goal was to determine the overall impact of the program compared with no program. The rest of the discussion would apply equally in this case.

impact could be estimated by comparing average prices (or number of jobs advertised) in LLMs assigned to each of the two program conditions. Estimating these impacts would require the collection of market-level outcome data.

Finally, although it is a clean and easily explainable design, the feasibility of randomly assigning at the level of the LLM may be problematic for two reasons. First, it may be difficult to concisely segment the overall market, i.e., define the LLM, as spillovers may occur at many different geographical levels. This problem is likely to be mitigated somewhat in studies of job search assistance programs designed to help low-wage workers, because the low-wage labor market is much more localized than the high-wage or professional market, and thus this may not be an especially difficult hurdle to overcome. Second, such a design requires the random assignment of a large number of clusters (LLMs) to achieve adequate statistical precision, with corresponding study organizational and data collection costs. We discuss this issue in greater detail in section 6.2.

### 5.2.2 Exploiting LLM-Level Nonrandom Variation

While ideally LLMs would be randomly assigned to different program groups, random assignment of entire LLMs to treatment and control (or to different program models) may not be feasible or desirable for the reasons just mentioned. This section outlines the options for estimating general equilibrium effects in those cases; first by using nonrandom variation in program models across sites, and then by using either nonrandom or random variation in the fraction of the population served by the program.

First, in some situations it is plausible that nonrandomized studies that compare across LLMs and are careful about causality issues may be able to achieve a similar result to cluster-randomized assignment, because the same general equilibrium effects will be captured by the aggregate LLM outcome measures used (Hsieh and Urquiola, 2006). Such an analysis will not usually be convincing, because local TANF agencies may select their JSA program model for reasons that are correlated with the outcomes of interest (e.g., employment success in a weak versus strong LLM). However, a situation where such a comparison is plausible might occur if, for example, a new program model were adopted in some LLMs but there was reason to believe that the decision to adopt the new program was made without regard to underlying characteristics of the LLM or program staff. Again, this is unlikely to be strictly true in practice—which is why random assignment is preferred—but may be sufficiently convincing after carefully controlling for observable characteristics of the LLM and program staff.

Another possibility for estimating general equilibrium effects is to utilize individual-level random assignment within LLMs in conjunction with LLM-level quasi-experiments exploiting variation in the fraction of participating job seekers between LLMs. This possibility is suggested by Ferracci, Jolivet, and van den Berg (2010), who describe a two-part design for measuring impacts of training programs on unemployed workers in France. A partial equilibrium effect is estimated in each LLM using an individual-level random assignment design, where some individuals are assigned to receive program services and the rest to a no-services control. (For the purpose of comparing two program models, we could layer a “new” program model—the “treatment”—on top of an existing program model—the “control”—in these sites). This treatment effect, by site, is then related to the fraction of treated workers in the LLM (compared with the entire estimated population of low-skill job seekers in the LLM) to see if the latter predicts average impacts for the LLM as a whole. If it does, the magnitude of the relationship suggests the size of the general equilibrium effect. In other words, if the estimated impact is smaller in sites where a higher fraction of job seekers were assigned to the treatment, the displacement effect can be

assumed to be large. If there is no relation between the fraction of treated job seekers and the estimated impact, displacement is inferred not to exist.

### 5.2.3 Double-Randomized Designs

A similar, but fully experimental approach would be to estimate the size of the equilibrium effect using a double-randomized design. In such a design, the specific individuals receiving the intervention would be randomly determined as above, but the fraction of individuals in each LLM receiving the tested intervention would also be randomly assigned. (Some LLMs could also be assigned to provide no JSA to TANF recipients, in order to estimate impacts relative to a “no program” counterfactual.) General equilibrium impacts would be estimated by the same method described in the previous subsection but with the advantage of being less subject to possible selection bias at the LLM level. According to Banerjee and Duflo (2008), Kremer and Muralidharan are using an approach like this to study the equilibrium effect of a vouchers program in villages in India. Crepon et al. (2012) have just reported findings from an application of this design to job placement assistance in France. A drawback of this type of design is that—like cluster-random assignment—it would need to incorporate a relatively large number of LLMs for statistical precision, a nonnegligible fraction of eligible job seekers in each LLM would need to receive the treatment, and the fraction of the job-seeking population who are program participants would have to be known.

### 5.2.4 Utilizing Nonintervention Local Labor Markets to Estimate Displacement

Another promising approach utilizing both experimental and nonexperimental data is outlined by Gautier et al. (2012), who compare intervention and nonintervention regions to analyze how externalities (i.e., violations of SUTVA) vary with treatment intensity. They study a Danish program for unemployed workers that was implemented at a small scale and in only some regions of the country—and within those regions individual-level random assignment was used to estimate the program impact. Administrative data on employment spells was available for individuals in both the intervention and nonintervention regions. Gautier et al.’s insight was to use outcome data from regions without the focal intervention to estimate both the positive impact of the program on participants and the *negative* impact on the control group, using the nonintervention areas as a counterfactual for both. Using a difference-in-differences model, they show that “nonparticipants in the [intervention] regions find jobs slower after the introduction of the program (relative to workers in other regions).” That is, by comparing the intervention and nonintervention regions, Gautier et al. can estimate the true (i.e., relative to a no-program counterfactual) treatment effect on individuals who get the intervention and the effect on those in the same LLM who do not get the intervention. Because the counterfactual regions are nonrandomly chosen, this design is nonexperimental. However, the difference-in-differences approach helps alleviate concerns about selection bias.

Gautier et al. then use these estimates to set the parameter values for a Diamond-Mortensen-Pissarides (Pissarides, 2000) equilibrium search model, which they use to simulate the effect of a large-scale rollout of the program. They show that although a standard partial equilibrium benefit-cost analysis would show positive results, such a rollout would actually decrease welfare due to congestion in the labor market. We are not interested in replicating this aspect of their design, because a potential TANF JSA evaluation would not expand the overall number of JSA participants. However, the basic approach could be used for comparison of one JSA intervention model to another model (or to no model) in a general equilibrium context.

Consider a situation in which an existing JSA model is widely implemented, and a new, more intensive model is being considered as an improvement. Following Gautier et al.'s strategy, a sample of LLMs could be selected to implement an individual-level experimental design to test the new model relative to the status quo. In these LLMs, the partial equilibrium impact would be estimated as the difference in outcomes between the new and status quo groups (who presumably collectively do not comprise the entire population of job seekers in that LLM). These outcomes would be compared with the average outcomes of LLMs that did not implement the experiment—and the total partial equilibrium impact would be divided between the (hopefully positive) effect on the new intervention group and the (likely negative) effect on the status quo group. The net impact—e.g., the net increase in employment due to the program differential—could be estimated as follows. First, sum the measured impacts on individuals in the intervention group to determine the total benefit to that population. Second, determine the total number of additional comparable job seekers in LLMs that implemented the experiment, including job seekers in the status quo JSA participant group, perhaps using Unemployment Insurance filings. Third, multiply the (likely negative) impact estimated for the status quo group by the total number of low-skill job seekers, to determine the total displacement for low-skill job seekers who do not participate in the new JSA intervention. Finally, subtract the total cost to job seekers not participating in the new JSA intervention from the total gain for job seekers who do participate in the new intervention to determine the net impact on the low-skill labor force—and therefore society—as a whole.

In contrast to a cluster-randomized design, this method does not require a large number of LLMs for identification of the equilibrium effects, limiting the cost of the evaluation. It does, however, require data on a large nonintervention sample of individuals. For cost purposes, it would be ideal to get this information from administrative sources (e.g., Unemployment Insurance records on employment and earnings).

### 5.2.5 Job Search Models Using Experimental Data Only

A final, promising approach to consider is the one described by Lise, Seitz, and Smith (2004), who also use an equilibrium job search model to estimate the general equilibrium effects of an earnings subsidy program called the Canadian Self Sufficiency Project (SSP). In contrast to Gautier et al., whose estimation strategy relies on differentials between regions, Lise, Seitz, and Smith use data only from regions with the intervention of interest to calibrate an equilibrium job search model, building on the work of Davidson and Woodbury (1993). The model is calibrated using experimental data on randomly assigned control groups of individuals in each site, to test the model's ability to replicate the observed behavior of each site's randomized treatment group. Once the model is properly calibrated and verified to correctly simulate treatment group outcomes, it can be recalibrated using data on the low-skilled population as a whole and used not only to estimate the displacement effect on the control group but to simulate the equilibrium effects that would result from introducing the policy in question as a general policy for all eligible individuals in the community.

The Lise, Seitz, and Smith approach does not require a large number of LLMs nor a large number of individuals receiving treatment in each LLM, but it does require difficult modeling techniques, and many additional assumptions are embedded in the model. The major advantage of this method is that, as the authors note, it is unusual for a policy to be implemented differently in separate jurisdictions of sufficient number to allow the econometric analyses outlined above. In the "usual" case then, a structural model will be required to estimate general equilibrium effects. Because the model used by Lise, Seitz, and Smith incorporates employers, it can be used to explore equilibrium effects on market-level variables like wages and prices as well as the typical employment outcomes.

### 5.3 Implications for DOSE

In sum, the following represent fully developed options in a general equilibrium analysis of worker displacement in a JSA impact evaluation:

1. Fully implement JSA (i.e., for all eligible individuals) in some LLMs but not in others, using cluster-randomized assignment. Estimate the equilibrium impact as the mean outcome in the treatment LLMs minus the mean outcome in the comparison (or control) LLMs. This strategy has the advantages that it is easy to explain and can be used for market-level outcomes (e.g., prices, wages in noncovered industries) as well as for individual outcomes (e.g., employment). Drawbacks are that the number of LLMs in the study would need to be fairly large, and that the LLM may be difficult to define.
2. Implement individual-level random assignment in some or all LLMs, but vary the proportion of individuals exposed to the treatment across LLMs. This strategy is easy to explain to a nontechnical audience, but may not adequately capture complex relationships between the fraction of individuals exposed to the treatment and the equilibrium effect of the program. This approach has many of the same drawbacks as cluster-randomized assignment.
3. Implement individual-level random assignment in a sample of LLMs, and collect data on outcomes in those and nonexperimental LLMs. Use this data to estimate the impacts on the treatment group and on the control group. With these estimates in hand, determine the total benefit to treated individuals and the cost to nontreated individuals to find the net impact. Compared with the preceding approach, this approach has the advantage that the proportion of individuals in the population assigned to treatment would not have to be systematically varied. In contrast to the previous two evaluation strategies this method does not require a large number of LLMs for identification of the equilibrium effects, limiting the cost of the evaluation. It does, however, require data on a large nonexperimental sample.
4. Implement individual-level random assignment for a subset of the target population. Use data on the control group to calibrate a Davidson and Woodbury (1993)-type equilibrium job search model. Use the model to estimate individual- and market-level outcomes from full implementation of the program in the relevant population. Advantages of this approach include the ability to estimate market-level equilibrium outcomes and the ability to account for complex relationships. The drawback is that a very complex model must be specified and estimated and strong assumptions made.

Any of these strategies could be expanded to include more than one treatment arm, or adapted to measure the impact of variations in individual JSA components.

Finally, we acknowledge that widescale adoption of a program could affect the program's takeup and utilization rates in the long term, which has general equilibrium implications. For example, a successful JSA program available only to TANF recipients could potentially increase the number of TANF applications and entries, while a JSA program that increased burden on recipients could decrease the number or change the composition of TANF applicants. Neither of these effects would be apparent in a typical partial equilibrium experiment.

## 5.4 Obtaining a General Equilibrium Impact Estimate from within the Basic Experiment

A final, original but not fully developed method, concerning spillover effects of a voluntary employment assistance program that randomizes individuals, concentrates *on the experimental sample* as the source of all impact information, including spillover effects on workers outside of the experimental sample displaced from jobs. As noted above, more intensive job search assistance may elevate some workers who would have been nonemployed with only low-intensity JSA into jobs and, commensurately, cause some of the workers who would have held those jobs to be nonemployed. This reshuffling of people and job “slots” complicates any experimental evaluation of JSA impacts involving randomly assigned individuals for reasons discussed above. But surprisingly, the experiment itself may offer a solution to this difficulty under certain conditions.

A model of the microeconomic process of job placements of specific workers in specific jobs develops this potential. Appendix A provides this model across several scenarios: with high- versus low-intensity JSA services and in both a random assignment context and in a “natural” environment. The analysis covers market-clearing outcomes with different intensities of JSA. It shows how *experimental* sorting yields treatment-control group comparisons that differ in important ways from the desired contrast between a world with high-intensity JSA and a world with low-intensity JSA (or two worlds with JSA programs that differ on other dimensions). This analysis points out the potential for experimental impact estimates to yield biased estimates of differential program effects, and simultaneously provides a potential means of obtaining unbiased estimates from the experiment alone.

Two distinct types of differences arise between the desired impact measure and the one provided by the experiment in this model, each producing a different type of bias in the standard experimental impact estimate:

- “Displaced-worker bias” caused by job losses among workers who are not part of the experimental sample, and
- “Extra-worker bias” caused by high-intensity JSA participants obtaining jobs they would not have obtained had the same services been delivered in the absence of a randomized experiment (see Appendix A for an explanation of this phenomenon).

“Program preferred random assignment,” suggested by Olsen, Bell, and Luallen (2007) in another context, offers a potential correction that removes both sources of bias. This strategy asks JSA program staff to identify TANF recipients most appropriate—or “preferred”—for high-intensity services prior to random assignment, allowing for separate impact estimation in the preferred and nonpreferred samples. Under certain conditions identified in the appendix, the separate estimates can be combined in a way that removes both bias sources from the overall impact estimate. While obviously an advantage, it is not clear how to ensure that the required conditions hold; more development work is needed to make the application of this technique more robust if possible—development work unfortunately beyond the scope of the presently provided funding.

## 6. Determine Needed Sample Sizes

The prior sections have presented design options for evaluating the impacts of JSA. Here we focus on the power of those designs to detect impacts of a particular size with a given sample size, which directly implies an evaluation's overall scale and cost dimensions—and also the value of the policy information it yields. For each design, we present minimum detectable effect sizes (MDESs) given reasonable assumptions, for pooled analyses and for a standard subgroup of particular interest. We then discuss these MDESs relative to plausible impacts and discuss implications for sample size.

The chapter first considers individual-level random assignment designs; it then considers designs that randomly assign sites. In the latter context, we consider the sample size implications of Chapter 5's options for measuring the general equilibrium effects of JSA. Given the MDES results, we revisit possible data collection strategies and their relative costs drawing from the earlier discussion of this topic from Chapter 3. We then close with thoughts about the possible implications of sample size needs for feasible designs.

### 6.1 Individual-Level Randomized Options

In Chapter 4, we discussed three individual-level randomized experimental designs: a basic randomized design with two contrasting treatments; a three-arm design; and a 2x2 factorial design. The sample size demands of these designs vary according to many factors. The minimum detectable impact of a given design is a function of the chosen power, level of statistical significance (including whether to conduct a one- or two-tailed test), the random assignment allocation ratio, and the variance associated with the outcome of interest. For example, measures that are “noisier”—they have higher variance—require larger samples to detect an effect of a given magnitude relative to measures with less noise. The ratio of the minimum detectable impact to the square root of the variance of the outcome measure (i.e., its standard deviation) is the “minimum detectable effect size,” or MDES, for a given study design. This tells us generically—i.e., for outcomes of all variance levels—how capable the design is of detecting impacts of policy consequence, since impacts in standard deviation units between 0.01 and 0.25 are thought to matter to policy in various contexts (though of course larger impacts would matter too).<sup>18</sup> These impact estimates are smaller than what one often considers to be a “small” effect size (by one common metric, Cohen considers an effect as small if it is 0.20 standard deviations and medium at 0.50). Our estimates are JSA-specific and based on prior research, which reveals that program effects are small. These effect sizes are commensurate with the relatively low cost of the intervention: small investment, small effect. The policy implication is clear: if a low-cost intervention can achieve even small effects, it might be a sensible investment.

Earnings tends to be a noisy measure, with its standard deviation (square root of variance) being as large as, and often larger than, its mean, whereas employment rates tend have less variability. These two

---

<sup>18</sup> In particular, impacts in standard deviation units—called effect sizes—indicate how far the intervention moves the average program participant up in the distribution of outcomes that would occur in the participant population absent the intervention. For example, an effect size of 0.10 moves a participant with an outcome at the median of the distribution (i.e., the 50th percentile) of the distribution of outcomes absent the intervention up to the 54th percentile with the intervention.

outcomes are candidates as a “confirmatory” outcome in a study of JSA programs’ effectiveness. We have identified earnings as the more central measure and recommend it as the primary, confirmatory outcome, as justified in Chapter 2. For that reason, we focus on earnings as the outcome of interest in computing MDESs here.

For each design, we first present MDESs for various sample sizes; then we present the sample size required to achieve a particular MDES. Training programs are more intensive than JSA programs. Given that a training program might be expected to generate an impact of 0.15 standard deviations, we propose that a reasonable effect size for a JSA versus no-JSA comparison would be 0.05 (standard deviation units). This corresponds to a \$105 increase in quarterly earnings or \$280 in annual earnings.<sup>19</sup> We base this information on the following evidence from, predominantly, the NEWWS, and The Los Angeles Jobs-First GAIN Evaluation. For example, the overall effect of the labor force attachment interventions studied in the NEWWS was about 0.12; the effect of the LA Jobs-First GAIN corresponds to roughly 0.089 standard deviations; and the differential effect of traditional vs. intensive job search was about 0.027 standard deviations.

In order to then think about the effect that one might observe between variants of JSA, we suggest that an effect size of 0.02 would be reasonable. It is also a relatively conservative figure, which in this situation is preferable. An MDES of 0.02 seems appropriately and sufficiently small that variants of a light-touch intervention like JSA could differ in impact by that amount. That said, we have focused our efforts on designs that make comparative assessments of variation in JSA programs. For such a comparison, we believe that 0.05 would be too large an effect to expect or detect. As a result, the threshold for determining samples sizes for this design effort should be the smaller effect size of 0.02, the more likely size of a differential impact of alternative treatments (rather than a treatment-control contrast). This corresponds to \$112 in annual earnings,<sup>20</sup> which seems a reasonably small difference to be able to detect, and a difference that variations in JSA service provision and/or content might reasonably be expected to produce.

Certainly a measure more sensitive (and less noisy) might make impact detection more likely, and so too might a treatment contrast that is greater than different variants of a light-touch intervention. To the extent that ACF would choose to implement a “maximum variation” experiment, in which very minimal JSA would be compared against a “pulling out all the stops” high-intensity version of JSA, an impact larger than 0.02 standard deviations is likely. For this possibility, we present the sample sizes needed to detect an effect of 0.04, should it exist. While still a relatively small effect size, it is double that of what our evidence would support as an expected differential effect size in contrasting JSA interventions that are not at the extremes of high and low intensity.

---

<sup>19</sup> In 2011 dollars, based on computations of standard deviations of earnings from the LFA sites in the NEWWS data. If we use JTPA as our source, a 0.05 effect size corresponds to \$208 in quarterly and \$830 in annual earnings for men and \$155 in quarterly and \$550 in annual earnings for women (also in 2011 dollars). Annual and quarterly earnings impacts vary in this manner because the standard deviation of annual and quarterly earnings differ relatively dramatically from one another. These estimates come from the NEWWS data as computed, adjusted and reported in Nisar, Juras, and Klerman (2012).

<sup>20</sup> Again, in 2011 dollars, based on computations of standard deviations from the NEWWS data.

Exhibit 6.1 shows the number of individuals who would need to be randomly assigned to attain minimum detectable effect sizes of various sizes in a two-arm individual-level randomization design with equal numbers assigned to each arm. A 0.02 MDES is achieved with 50,000 cases and a 0.04 MDES with around 12,500.

We explain some assumptions underlying these computations. First, we use the conventional power level of 80 percent (though certainly 70 or 90 could be justified). We use a two-tailed test because some prior research reveals evidence of negative effects of training on some subgroups over some time periods. A unidirectional hypothesis is therefore not appropriate. Further, we choose to use an alpha of 0.10. The main alternative to a 10 percent significance level—an alpha of 5 percent—would tilt a confirmatory hypothesis test too much in the direction of avoid false positives (Type I error) while unduly expanding the risk of finding false negatives (Type II error: that is, finding no significant impact when an effect of important magnitude in fact occurs). Combined with the recommended added conservativeness of imposing an adjustment for conducting multiple hypothesis tests, we feel justified in using an alpha of 0.10 initially here. Next, based on prior studies, it seems reasonable to assume that including regressors (e.g., demographics, lagged earnings) will lead to an R-squared of roughly 20 percent. Finally, we assume that the ratio of assignment to alternative treatment arms is 1:1. The ratio in any given design, in practice, might veer from this ratio, in which case MDESs would be larger than what we estimate here—and to achieve a given MDES, samples would need to be larger.

**Exhibit 6.1: Minimum Detectable Effect Sizes (MDESs) for a Two-Arm Design with Individual-Level Random Assignment, by Sample Size**

| Sample Size | MDES for T2-T1 |
|-------------|----------------|
| 2,500       | 0.089          |
| 5,000       | 0.063          |
| 7,500       | 0.051          |
| 10,000      | 0.044          |
| 20,000      | 0.031          |
| 30,000      | 0.026          |
| 40,000      | 0.022          |
| 50,000      | 0.020          |
| 60,000      | 0.018          |
| 70,000      | 0.017          |
| 80,000      | 0.016          |
| 90,000      | 0.015          |
| 100,000     | 0.014          |
| 110,000     | 0.013          |
| 120,000     | 0.013          |

Notes: These calculations assume a 10% level of significance for a two-tailed test, 80% power, an R-squared of 0.20 and 50% of individuals assigned to T1.

If instead a three-arm design with individual randomization was adopted (again with equal cell sizes), Exhibit 6.2 suggests that a 0.02 MDES would require a total sample size of around 75,000 cases and a 0.04 MDES a total sample size of just under 20,000.

**Exhibit 6.2: Minimum Detectable Effect Sizes (MDESs) for a Three-Arm Design with Individual-Level Random Assignment, by Sample Size**

|         | MDES for T1-C | MDES for T2-C | MDES for T2-T1 |
|---------|---------------|---------------|----------------|
| 2,500   | 0.109         | 0.109         | 0.109          |
| 5,000   | 0.077         | 0.077         | 0.077          |
| 7,500   | 0.063         | 0.063         | 0.063          |
| 10,000  | 0.054         | 0.054         | 0.054          |
| 20,000  | 0.039         | 0.039         | 0.039          |
| 30,000  | 0.031         | 0.031         | 0.031          |
| 40,000  | 0.027         | 0.027         | 0.027          |
| 50,000  | 0.024         | 0.024         | 0.024          |
| 60,000  | 0.022         | 0.022         | 0.022          |
| 70,000  | 0.021         | 0.021         | 0.021          |
| 80,000  | 0.019         | 0.019         | 0.019          |
| 90,000  | 0.018         | 0.018         | 0.018          |
| 100,000 | 0.017         | 0.017         | 0.017          |
| 110,000 | 0.016         | 0.016         | 0.016          |
| 120,000 | 0.016         | 0.016         | 0.016          |

*Notes:* These calculations assume a 10% level of significance for a two-tailed test, 80% power, an R-squared of 0.20, one-third of individuals assigned to T1 and one-third assigned to T2. In this case, we consider “C” to represent a basic JSA intervention, with T1 and T2 representing selected variants of interest to evaluate.

A 2x2 randomized factorial design has still different statistical properties, as shown in Exhibit 6.3. Similar to a simple two-arm design, the comparison of any one factor to the status quo, shown in the first column, requires 50,000 cases in total for an MDES of 0.02 and around 12,500 for an MDES of 0.04. More total sample is needed to achieve the same level of statistical precision for other types of factor comparisons as illustrated in the remaining columns of the exhibit, including 100,000 cases for a 0.02 MDES and around 25,000 cases for a 0.04 MDES. It is worth noting that the full exploration of status quo versus Factor A versus Factor B versus Factors A-plus-B in a conventional four-arm design would not yield such low MDESs for the comparison of any one factor to the status quo as shown here; rather, it would face a sample size/MDES tradeoff equivalent to that in the other three columns. As noted earlier, the added power in a factorial design comes from the use of all of the non-Factor-A-assigned individuals as the control group for Factor A, and the non-Factor-A-assigned group includes those assigned to the status quo as well as to Factor B.

**Exhibit 6.3: Minimum Detectable Effect Sizes (MDESs) for an Individual-Level 2x2 Randomized Factorial Design, by Sample Size**

| Sample Size | MDES  | MDES for Factors | Differential MDES | Differential MDES |
|-------------|-------|------------------|-------------------|-------------------|
| 2,500       | 0.089 | 0.126            | 0.126             | 0.126             |
| 5,000       | 0.063 | 0.089            | 0.089             | 0.089             |
| 7,500       | 0.051 | 0.073            | 0.073             | 0.073             |
| 10,000      | 0.044 | 0.063            | 0.063             | 0.063             |
| 20,000      | 0.031 | 0.044            | 0.044             | 0.044             |
| 30,000      | 0.026 | 0.036            | 0.036             | 0.036             |
| 40,000      | 0.022 | 0.031            | 0.031             | 0.031             |
| 50,000      | 0.020 | 0.028            | 0.028             | 0.028             |
| 60,000      | 0.018 | 0.026            | 0.026             | 0.026             |
| 70,000      | 0.017 | 0.024            | 0.024             | 0.024             |
| 80,000      | 0.016 | 0.022            | 0.022             | 0.022             |
| 90,000      | 0.015 | 0.021            | 0.021             | 0.021             |
| 100,000     | 0.014 | 0.020            | 0.020             | 0.020             |
| 110,000     | 0.013 | 0.019            | 0.019             | 0.019             |
| 120,000     | 0.013 | 0.018            | 0.018             | 0.018             |

Notes: These calculations assume a 10% level of significance for a two-tailed test, 80% power and an R-squared of 0.20. We further assume that a quarter of the sample receives T1 only, a quarter receives T2 only, a quarter receives T1 and T2 and the remaining sample receives the base treatment (in this case the “control” condition). The difference between the first two MDES columns is that the first one uses anyone not randomized to the treatment arm as a control group (so T2 and C together provide the counterfactual for T1, for example); whereas the second one uses the control (base treatment) group as the counterfactual for the combined effect.

Precise sample size targets for a 0.02 MDES and a 0.04 MDES differ slightly from the approximate numbers suggested by this analysis of MDES ranges. Exhibit 6.4 contains that information, providing an all-in-one-place reference for sample size requirements under different target scenarios across varied individual-level randomized designs.

**Exhibit 6.4: Sample Size Requirements for Minimum Detectable Effect Sizes (MDESs) of 0.02 and 0.04, by Design Option with Individual-Level Randomization**

| Design Option              | MDES of 0.02 | MDES of 0.04 |
|----------------------------|--------------|--------------|
| Basic (Two-Arm) Randomized | 46,708       | 12,366       |
| Three-Arm                  | 74,121       | 18,548       |
| Factorial                  | 98,921       | 24,731       |

Notes: These calculations assume a 10% level of significance for a two-tailed test, 80% power and an R-squared of 0.20. See tables above for descriptions of how the sample is allocated to treatment conditions.

In terms of expected subgroup impacts, we know that, for a subgroup of half the size of the overall sample, the smallest detectable impact would be about 1.4 times that of the overall impact. This relationship emerges in Exhibit 6.5, which shows subgroup MDESs by the size of the subgroup involved as a proportion of the overall sample. With a target of 0.02 differential effect size, we get a subgroup MDES of 0.028 for a subgroup consisting of half the sample. The smaller sample size that comes with subgroup analyses is generally thought to weaken a design’s ability to detect small impacts; and it certainly does so if we demand that the MDES is held constant. That is, if detecting the same differential impact for a subgroup would be of interest, then sample sizes would need to increase to do so. Generally, however, it is of less interest to know whether selected subgroups experience the average treatment effect; instead, it is of interest to know whether selected subgroups experience a larger or smaller impact. Because the sample sizes are already quite large, we do not suggest powering the design to detect

subgroup effects that would be smaller than we would expect to detect for the sample overall. Instead, we believe it is appropriate to devise subgroup analyses to detect somewhat larger impacts than with estimating the overall, average treatment effect. With this approach, given sample sizes do not necessarily become hopelessly large for assessing subgroup effects.

The proportion of the sample that falls within a likely subgroup of interest determines the loss of power associated with restricting the analysis to the subsample. Focusing on individuals with low baseline education or individuals with weak past labor market attachment, we assume that each subgroup will make up approximately 50 percent of the overall sample. Using this rough guideline is justified from existing data (although the exact current proportion might differ from these past estimates). For example, the proportion of individuals who have no high school diploma and no GED is 44 percent in Bloom, Hill and Riccio’s (2003) pooled data from several welfare-to-work program evaluations and 47 percent in Parisi et al.’s (2006) analysis of administrative data on TANF recipients. In Bloom, Hill and Riccio’s (2003) pooled sample, 56 percent had zero earnings in the previous years.

Subgroups comprising a smaller share of the total sample would have larger MDESs, and subgroups comprising a larger share of the total sample would have smaller MDESs. The table below gives the subgroup MDES for a variety of alternative subgroup proportions. This table answers the question: what would the subgroup MDES be if we power the study to detect a differential impact of 0.02. The answer to the question is independent of the chosen design because the subgroup MDES is mathematically determined by the overall MDES and the subgroup proportion. The subgroup MDESs in the table below might be achieved using a basic randomized design with a sample size of 7,914, a three-arm design with a sample size of 11,871, or a factorial design with a sample size of 7,914.

**Exhibit 6.5: Subgroup Minimum Detectable Effect Sizes (MDESs) for a Design with a 0.02 Overall MDES, by Size of Subgroup**

| Subgroup Proportion | Subgroup MDES |
|---------------------|---------------|
| 10%                 | 0.063         |
| 20%                 | 0.045         |
| 30%                 | 0.037         |
| 40%                 | 0.032         |
| 50%                 | 0.028         |

*Notes:* These calculations assume a 10% level of significance for a two-tailed test, 80% power and an R-squared of 0.20. The full-sample MDES is 0.02; these calculations use the specified subgroup proportion.

## 6.2 Cluster-Randomized Options

As noted in Chapter 4, cluster-randomized designs are appropriate when individuals cannot be randomized or one expects spillover effects between random assignment arms with individual-level assignment of workers in the same local labor market. Determining adequate sample sizes when randomizing aggregate units involves an additional assumption about the cross-cluster variation in the key outcome. When cross-site variation in average characteristics is small, fewer sites will be required to deliver adequate statistical precision than when cross-site variation is large. Within-site versus between-site variability is captured in the so-called “intra-class correlation” (ICC), defined as the proportion of total outcome variance that exists across rather than within sites. ICCs are commonly computed in educational settings, where clustering can play a big role in powering evaluation studies. Clustering is somewhat less of a concern—but still of some concern—in welfare and employment and training research, where relatively less across-cluster variability exists.

In addition to our assumptions detailed earlier (re: power, two-tailed test, alpha), these cluster-randomized designs' MDESs rest on three more assumptions related to this cross-cluster variability as noted next (re: ICC, R-squared, and number of observations per site). We use the most recent estimates of ICCs in welfare and employment and training settings from Klerman, Juras, and Nisar (2012) to generate our computations for the sample sizes needed to power a cluster-randomized experiment to evaluate the effectiveness of JSA programs. In that work, the computed ICC ranges from 0.016 for quarterly earnings, 0.024 for annual earnings, and 0.033 for total five-year earnings from the NEWWS data.<sup>21</sup> We use an ICC assumption of 0.03 here; see Exhibits 6.6 and 6.7 for MDES information with various cluster-level random assignment designs. Klerman, Juras and Nisar (2012) also compute site-level R-squareds ranging from 0.217 to 0.398, so we assume a cluster-level R-squared of 0.25. Finally, they compute individual-level R-squareds ranging from 0.087 to 0.155; so we assume an individual-level R-squared of 0.10. These computations also rely, to a certain degree, on the average number of individuals per site. We assume that each site will have 100 individuals, which reflects an upper bound on power (smaller samples per site mean that the ICCs will have a relatively larger influence) and our assumption about a realistic sample size per site to target with this sort of design. As such, our numbers conveniently mirror the individual-level MDES options reported by sample size: a 100-site design might be considered to correspond to a 10,000-individual design, and we can thereby easily make comparisons in the MDESs between the individual- and cluster-level randomization.

**Exhibit 6.6: Minimum Detectable Effect Sizes (MDESs) for Site-Level Cluster-Randomized Design, by Number of Sites**

| Number of Sites | MDES for T2-T1 |
|-----------------|----------------|
| 50              | 0.124          |
| 100             | 0.088          |
| 200             | 0.062          |
| 300             | 0.051          |
| 400             | 0.044          |
| 500             | 0.039          |
| 600             | 0.036          |
| 700             | 0.033          |
| 800             | 0.031          |
| 900             | 0.029          |
| 1,000           | 0.028          |

Notes: These calculations assume a 10% level of significance for a two-tailed test, 80% power, an ICC of 0.03, a site-level R-squared of 0.25, an individual-level R-squared of 0.10, 100 individuals per site and 50% of individuals assigned to treatment.

<sup>21</sup> The corresponding ICC from the JTPA data range from 0.016 to 0.047, depending on the populations covered and the specific earnings measure used.

**Exhibit 6.7: Minimum Detectable Effect Sizes (MDESs) for Double Randomized Design, by Number of Sites**

| Number of Sites | MDES for T1-C | MESI for T2-C | MDES for T2-T1 |
|-----------------|---------------|---------------|----------------|
| 50              | 0.094         | 0.094         | 0.116          |
| 100             | 0.067         | 0.067         | 0.082          |
| 200             | 0.047         | 0.047         | 0.058          |
| 300             | 0.039         | 0.039         | 0.047          |
| 400             | 0.033         | 0.033         | 0.041          |
| 500             | 0.030         | 0.030         | 0.037          |
| 600             | 0.027         | 0.027         | 0.033          |
| 700             | 0.025         | 0.025         | 0.031          |
| 800             | 0.024         | 0.024         | 0.029          |
| 900             | 0.022         | 0.022         | 0.027          |
| 1,000           | 0.021         | 0.021         | 0.026          |

*Notes:* These calculations assume a 10% level of significance for a two-tailed test, 80% power, an ICC of 0.03, a site-level R-squared of 0.25, an individual-level R-squared of 0.10, 100 individuals per site, one-half of sites assigned to T1 and the other half to T1, and half of individuals within each site assigned to the treatment condition and the remaining to the status quo control condition.

An MDES of 0.02 requires even more sites than shown in either of these tables. Exhibit 6.8 contains the exact site counts needed to reach this level of statistical precision, including a previously unexplored three-arm site-level randomization design. It provides an all-in-one-place reference for sample size requirements for three different site-level random assignment designs.

**Exhibit 6.8: Number of Sites Needed to Achieve Target MDES of 0.02 by Design Option, with Site Randomization**

| Design            | MDES of 0.02 |
|-------------------|--------------|
| Basic Randomized  | 1,824        |
| Multi-Arm         | 2,897        |
| Double-Randomized | 1,113        |

*Notes:* These calculations assume a 10% level of significance for a two-tailed test, 80% power, an ICC of 0.03, a site-level R-squared of 0.25, an individual-level R-squared of 0.10 and 100 individuals per site. See tables above for descriptions of how the sample is allocated to treatment and control conditions.

We do not examine subgroup impacts within the context of site-level randomized designs because the initial examination of site-level sample sizes needed identifies a sample size that is prohibitive of using the design option. To extend further the analysis to consider subgroup effects would only exacerbate the situation.

**6.3 General Equilibrium Effect Analysis**

Chapter 5 identifies two methods as having potential for effectively estimating the general equilibrium effects of JSA:

- Cluster-randomized assignment of entire local labor markets (LLMs); and
- “Program preferred” random assignment of individuals within LLMs.

The former design has the same MDESs as cluster designs in general, as explored in the previous section. The difference is that true effects may be slightly larger—and hence more detectable for any sample size—when effects on labor market outcomes can occur through indirect channels. However, we would not expect the general equilibrium spillover effects to be large relative to direct effects on JSA participants, so our appraisal of the adequacy of cluster-level sample sizes from above does not change.

The use of “program preferred” (PP) random assignment and concurrent analysis of “next tier applicants” (NTAs) to measure general equilibrium effects within the experimental sample does have MDES and sample size implications. From Chapter 5’s Appendix A, an unbiased measure of average impact that takes account of general equilibrium effects by this means is

$$\text{Impact}_{\text{NTA}} - (\frac{1}{2}) \text{Impact}_{\text{PP}}.$$

It can be shown that this estimator has a standard error and MDES 70 percent larger than a simple treatment-control difference based on the same total sample size, assuming a 2:1 treatment-to-control-group ratio for the PP group and a 1:2 ratio for the NTA group. Thus, the ability to detect statistically significant impacts from this analysis is likely lower than appraised above, since general equilibrium effects are not likely to add 70 percent to true impacts beyond direct effects on JSA participants.

## 6.4 Data Collection Strategies Revisited

Chapter 3 discussed broad strategies for data collection. The discussion here in Chapter 6 has emphasized that estimating the impact of JSA is likely to require large samples, such that evaluation costs—and, in particular, data collection costs—are likely to be major considerations in design. Therefore, this section reconsiders the potential data collection strategies originally introduced in Chapter 3.

The logic model developed in Chapter 2 suggested that the most likely impacts of JSA are in the short term, i.e., getting the participant a job faster—during the 12 weeks of JSA versus several months later, or even at the next round of allowed JSA in the next year. Results for LA Enhanced Job Club (Navarro, Azurdia, and Hamilton, 2008) are consistent with this conjecture (i.e., impacts are concentrated almost exclusively in quarters 2 and 3 after randomization), as are the results of the DOL UI experiments (see Klerman et al., 2012). As noted in Chapter 2, longer-term effects (e.g., 36 months) on wages and other job characteristics—positive or negative—are theoretically indicated, but seem likely to be smaller than the short-term effects.

This range of outcomes suggests several different data collection strategies. In as much as quarterly earnings (or quarterly employment) is the sole outcome of interest, administrative data (i.e., UI earnings, probably as collected in the NDNH) are sufficient; no survey is necessary. Furthermore, such UI earnings data are relatively inexpensive, so both short-term and long-term exploration of effects on earnings (but not other job characteristics) would be possible at relatively low cost. On the other hand, insofar as time to first employment is an important outcome, UI earnings data are insufficient except to measure whether the first paid (and UI-covered) job occurs in the first calendar quarter after program entry/random assignment or the second calendar quarter (or neither). Instead, a survey that could document finer-grained timing of employment, either at 6 or 12 months after random assignment (assuming a reliable history could be constructed to reveal details about the timing of employment) or even more frequently, seems necessary.

It is also possible that time to first employment will demonstrate a larger effect size than earnings and could be examined based on a survey that, for cost savings, includes just a random subsample of the full evaluation sample and thus a smaller sample size. Given these considerations, three strategies seem reasonable:

1. **Data-Intensive Strategy.** Survey the entire study sample at 6 or 12 months, and collect UI earnings data (probably through NDNH) at least through 12 months and probably through 36 months. This strategy would allow equal focus on quarterly earnings and time to employment.<sup>22</sup>
2. **Pure Administrative Data Strategy.** Collect no followup survey data. Instead, only use UI earnings data. This would be the lowest cost strategy, but it would preclude estimating impacts on time to employment except in the crude sense of whether the contrasting interventions lead to different shares of the population holding some paid job in the first quarter after random assignment or in the second quarter after random assignment.
3. **Small Survey Strategy.** A mixed strategy might proceed with UI earnings data and also a smaller survey of, perhaps, half of the sample. Such a smaller survey would be consistent with the available evidence that time to first employment is a more sensitive indicator, and therefore impacts could be detected with a smaller sample. This strategy would be intermediate in cost.

These three options consider only the nature of the followup survey. They also assume a conventional telephone survey (presumably with in-person followup), as is common in this type of evaluation. If a followup survey can be markedly less expensive than the conventional approach (as might be the case with a web- or text-based survey), then the strength of the recommendation behind these options would be different. We expect that any design would include a baseline survey at sample entry, and UI earnings data through 12 (and probably 36) months, as well as potentially process and benefit-cost analyses as discussed in Chapter 7. Alternative possibilities exist for data collection that would demand further reconsideration of the strategies suggested here. Little evidence exists to provide a rationale for embarking on a web- or text-based survey in this context, but if ACF were interested in forging new ground, exciting opportunities exist. Particularly with respect to the text-survey idea, advance pilot testing should occur, given the paucity of information from existing research on text survey response rates for this target population (or any, for that matter). Further, given large potential sample sizes, additional learning about the differential effectiveness of various modes of data collection could occur (that is, experimenting on modes of data collection within a JSA evaluation).

Exhibit 6.9 presents some rough estimates of cost associated with two modes of data collection (conventional telephone and a text-based approach) and tracking of study participants needed regardless of the mode. It makes clear that a large telephone survey can be very expensive. While the estimate of a small (n=5,000) weekly text-based survey is about half the cost of a 15-minute telephone survey, the marginal cost of more sample in text mode is markedly less than in telephone mode. With a sample of 75,000, the weekly text-based survey is one-tenth the cost of a 15-minute followup survey. A clear tradeoff here is that the questions about the timing of employment and amount of earnings must be extremely brief to be asked by text. To the extent that they could be validated, the benefit is that the increased frequency of observation in text-based mode is in line with ACF's interest in capturing variation in the timing of employment. Again, regardless of the mode of followup (be it survey by telephone, text or some other mode, or administrative data), we assume a baseline survey would be a key part of the evaluation research.

---

<sup>22</sup> If time to first employment is also designated as a confirmatory outcome along with earnings, a multiple comparison adjustment would require a sample size about 22 percent larger in order to detect the same true effect size on one or both confirmatory outcomes.

**Exhibit 6.9: Followup Survey Cost Estimates for Varied Options**

| Survey Length →<br>Mode, Sample Size | 6-Month Followup |           | 12-Month Followup |           | 6+12-Month Followup |            |
|--------------------------------------|------------------|-----------|-------------------|-----------|---------------------|------------|
|                                      | 15 min           | 30 min    | 15 min            | 30 min    | 15 min              | 30 min     |
| Phone Survey                         |                  |           |                   |           |                     |            |
| 5K                                   | 442,200          | 604,300   | 435,700           | 583,500   | 877,800             | 1,187,700  |
| 15K                                  | 937,800          | 1,404,400 | 910,800           | 1,351,300 | 1,848,500           | 2,755,600  |
| 30K                                  | 1,681,100        | 2,604,400 | 1,623,600         | 2,503,000 | 3,304,700           | 5,107,300  |
| 50K                                  | 2,734,200        | 4,266,300 | 2,637,700         | 4,102,500 | 5,371,800           | 8,368,700  |
| 75K                                  | 3,973,200        | 6,266,300 | 3,825,600         | 6,022,100 | 7,798,800           | 12,288,400 |
| Weekly for 12 months                 |                  |           |                   |           |                     |            |
| Text Survey                          |                  |           |                   |           |                     |            |
| 5K                                   |                  |           | 213,300           |           |                     |            |
| 15K                                  |                  |           | 237,300           |           |                     |            |
| 30K                                  |                  |           | 273,300           |           |                     |            |
| 50K                                  |                  |           | 321,300           |           |                     |            |
| 75K                                  |                  |           | 381,300           |           |                     |            |
|                                      | 6-month alone    |           | 12-month alone    |           | 6+12-month both     |            |
| Tracking                             |                  |           |                   |           |                     |            |
| 5K                                   | 29,100           |           | 27,200            |           | 56,200              |            |
| 15K                                  | 77,400           |           | 76,400            |           | 153,800             |            |
| 30K                                  | 149,600          |           | 150,200           |           | 299,800             |            |
| 50K                                  | 245,900          |           | 248,500           |           | 494,400             |            |
| 75K                                  | 366,200          |           | 371,300           |           | 737,500             |            |

Notes: These costs do not represent the exact costs of any given followup strategy nor do they constitute an actual proposal to implement the survey. Instead, these are rough estimates based on many assumptions, available on request. Estimates cover implementation of survey and tracking efforts only and do not include other aspects of instrument development, site and/or participant recruitment, baseline data collection, in-person followup (as may be needed to ensure high response rates) or incentive payments to respondents (also as may be needed to ensure high response rates). Therefore these estimates *underestimate* cost to implement a given option, but we hope provide a sense of the relative costs across varied options, by sample size, survey mode/length, and followup period/frequency.

Given these data collection options as motivated by sample size considerations, the crucial issue for ACF concerns the interplay between available budget and the priority the government places on time to first employment as an area of impact relative to impacts on earnings. We further revisit this topic in the concluding chapter of this report.

## 6.5 Discussion and Conclusion

In this chapter, we discuss and interpret the MDESs in light of the smallest true effects that (a) it is reasonable to expect and (b) that are important to detect for policy purposes, in order to identify the particular sample sizes that are sufficient to achieve each of the evaluation design's goals. We have pegged the evaluation's target MDES at 0.02 *for a differential effect*, which we think is justified given that an overall treatment effect of JSA versus no-JSA would likely generate an effect of about 0.05. In a situation where a contrast greater than that between two variants of a light-touch intervention would be examined, there is justification for considering a larger MDES. We also suggest that an appropriate MDES for subgroups, on which we might expect to observe larger effects, might be somewhat larger still, thereby not posing a restrictive sample size penalty in any of the possible designs. Three final points are relevant in concluding.

First, Chapter 4 noted that, regardless of whether the evaluation would target sites or individuals as the units of randomization, the evaluation must involve multiple sites. At the very least, sites might be selected to provide sufficient individual-level sample for the impact analysis, although they might also be

systematically selected to provide desired variation in contextual characteristics along the dimensions of the conceptual framework on which diversity is helpful: state TANF policy parameters, job seekers' characteristics, and local labor market conditions. Moreover, sites might be selected randomly, either in support of an individual-level randomized design or as part of a site-level randomized design, where the latter would require a larger number of sites than an individual-level randomization would demand. The distinction, therefore, across various site-level strategies has relatively few implications for sample size. That is, many sites will be needed regardless.

Next, and related to the prior point, is our conclusion that a group-randomized design seems prohibitive in terms of site-level sample sizes: 300 or more sites would be needed to detect an effect of 0.05 standard deviations in earnings; more than 1,000 sites in order to detect an effect of 0.02. With a sample this large, certainly the other contextual considerations could be well met, but the recruitment and monitoring effort seems particularly daunting. It is our assertion, therefore, that sample sizes dictate focus on a design in which individual-level randomization occurs. Practical issues of recruitment aside (though a topic to which we return in Chapter 8), the number of sites needed to support intake of 50,000–75,000 or more sample members is still large, perhaps as many as 50 in number. Despite the likely challenges that would accompany such a recruitment effort, we believe that securing that many sites in the evaluation of JSA would likely provide the contextual variation needed to explore the role of local context—in policy, population and labor markets—that are key questions for this research.

Finally, the minimum detectable effect size that we choose to power for is small (perhaps \$100 per quarter) and required sample sizes are large because changing only a small part of the JSA program is unlikely to imply major changes in impact. To make the issues clear, consider the largest possible contrast. One JSA program might involve minimal JSA intensity (e.g., an initial lecture to explain expected job search and how it will be monitored, followed by verification at weeks 3 and 6) versus an intensity of staff time for JSA nearly twice the current level. (These two interventions are deliberately specified such that the net impact is cost neutral in staff time.) Contrasts of this form are policy relevant. If there are not large differential impacts, then states could sharply reduce expenditures on JSA programs. Furthermore, it is plausible that differential impacts of these two approaches might be on the order of \$325. In as much as the target impact is \$325, the implied samples are a tenth the size. Clearly, the nature of the intervention is crucial for design, cost, and feasibility. We return to this issue in Chapter 8.

## 7. Evaluation Components to Complement a JSA Impact Study

A well-designed and successfully executed impact evaluation comparing two or more JSA interventions for TANF recipients will tell policymakers how much more those recipients benefit from one of the interventions than the other(s). It will not tell them:

- What the JSA program consists of, how the alternative interventions were implemented, and what aspects of the process of JSA service delivery distinguish the intervention models or components: the purview of a *process analysis*.
- What factors are associated with some TANF recipients participating in a particular JSA intervention in the first place or taking a particular course through the various elements that the intervention offers: the purview of a *participation analysis*.
- The extent to which the added benefits for participants in the more successful intervention translate into better outcomes for society as a whole, once the resource costs of the JSA programs are compared and possible consequences of the interventions for segments of society—including government (taxpayers)—beyond participants are weighted in the balance: the province of *benefit-cost analysis*.

The DOSE design team does not intend to produce a blueprint for any of these three evaluations components at the level of detail it has done for the impact analysis in previous chapters. However, ACF has asked the design team to consider how the ability of the government to understand and interpret the impact findings would be enhanced were further evaluation components designed and added to an impact study as a larger evaluation “package” for assessing the contribution of the various JSA approaches evaluated.

This chapter discusses how these three components would enhance an impact analysis. Specifically, the chapter begins by explaining how process, participation, and benefit-cost analyses would complement the impact analysis and make its findings more valuable for policy decision-making. We then lay out the broad parameters for each of these complementary evaluation components and their data requirements. Considering the costs of added data collection and analysis required for the added components, we conclude by recommending that ACF expand the design and implementation efforts for a future JSA evaluation—if undertaken—into all three added areas.

### 7.1 Process Analysis

A process analysis will contribute substantially to a straightforward and detailed understanding of how each of the participating sites actually operates the JSA intervention by: (1) describing the JSA models’ design and operation in each site, concentrating on measuring *differences* between the contrasted interventions as implemented; (2) identifying specific lessons for improving policy and practice and allowing replication in other locations; and (3) helping interpret outcome results. More specifically, it would:

- Document and contrast the exact nature of the treatments delivered;
- Document how each treatment was implemented and the context in which it operated;

- Identify challenges encountered in the process of generating each treatment and if and how they were resolved;
- Determine the extent to which interventions were implemented as planned (were goals such as effective collaboration among participating agencies, achievement of intake targets, and delivery of services on a timely basis achieved?);
- Document successes and challenges of the interventions examined and the inputs used to generate each intervention (e.g., organizational structure, operating procedures, community involvement, funding); and
- Provide context to help interpret effects on participants (including why certain impacts are detected, or not, as related to operational factors).

Addressing these topics would require several design, data collection, and analysis/reporting steps in an accompanying process analysis:

1. Choose the data collection approach for each of the above topics.
2. Determine the samples/units to be covered by each data collection component.
3. Develop analytic procedures for addressing each of the research topics from those sources.
4. Determine how process study findings can be integrated with those of the other included evaluation components—particularly the impact analysis findings (but also participation analysis and benefit-cost analysis findings, if any)—to enhance the learning and policy guidance achieved by the evaluation agenda as a whole.
5. Collect the needed quantitative data through appropriate data collection modes, such as online surveys of JSA providers and/or major employers, management information system (MIS) data from providers on services delivered to individual clients, published information about local economic conditions and labor market aggregates, and surveys of individual participants in the various JSA interventions to be compared in the impact analysis.
6. Collect the needed qualitative data through a combination of telephone interviews or surveys, and site visits that include several data collection approaches such as in-depth interviews with JSA provider agency management and line staff and employers, observation of delivery of JSA services, review of case records and agency documents on policies and procedures, and focus groups with clients (drawing on an existing battery of constructs used in other HHS-funded research, such as in GAIN or NEWWS).
7. Conduct the planned analyses and report the results.
8. Relate the process study findings to the impact analysis results and findings from any other evaluation components to determine, in particular, what aspects of JSA service delivery distinguish the interventions being compared in the impact and benefit-cost analyses and how the variants of JSA might be strengthened operationally in the future.

As to scale, one might consider doing a full process analysis at each site. However, this would involve a major resource investment, given the set of steps and data sources listed above and the large number of

sites needed for the impact analysis.<sup>23</sup> Alternatively, one might consider representing the full collection of impact analysis sites as much as possible through some type of systematic (or even random) sampling of process study sites, at least for conventional process study elements requiring on-site, time-intensive data collection activities and descriptive analysis of large volumes of qualitative information. In addition to, or in lieu of, selecting a subset of sites for inclusion in the process study, to make the number of visited sites manageable, alternative data collection and analysis approaches could be used that focus on remotely administered surveys, extraction of existing MIS data from agency computer systems, and summative quantitative analysis of these data covering all sites.

In this vein, telephone interviews with program managers and line staff in each site might allow researchers to document decisions around the design, implementation, and ongoing operations of the intervention at less cost than on-site data collection. More detailed site visits to a subset of the studied programs would then allow researchers to gather more detail about program policies, procedures, and operations as well as client experiences, where these topics seem likely to be most informative or salient.

A critical responsibility of a contractor hired to design a process study of JSA—should ACF choose to pursue that course in moving beyond the current project—will be to determine how broad the process study’s coverage needs to be for each research topic addressed. And, where it cannot be universal, to propose where to focus data collection and analysis among the various topics listed above and the many sites likely to form the basis for the impact analysis. If chosen well, this complementary research will certainly enhance ACF’s understanding of how JSA programs for TANF recipients evolve in the field and might be strengthened operationally in the future.

## 7.2 Participation Analysis

Two major factors conditioning the relative impacts of different JSA interventions are:

- Who participates in the two interventions contrasted in the impact analysis, and
- What paths participants follow through the potential elements of the various JSA strategies examined.

If researchers can tell policymakers which TANF recipients participate—i.e., comply with JSA participation requirements—under the contrasting approaches and, of those who participate, which recipients engage in which types of JSA activities, the evaluation will illuminate potential areas for targeting and possible refinements to the outreach/intake/service delivery approaches of the tested strategies. Specific participation questions of interest include:

- To what extent do individuals targeted for JSA participation participate, as opposed to not responding to their participation mandate?
- What characteristics distinguish participants from nonparticipants?

---

<sup>23</sup> Recall that prior chapters established that a multisite study of many JSA providers and local labor markets will be needed, likely in the dozens.

- To what extent do individuals in each intervention group use the particular individual services provided by their JSA intervention program and to what extent to they do so indistinctive ways between the two intervention models contrasted in the experiment?
- Who uses each major type of service, and how does this differ across variants of the JSA intervention?
- Do participation patterns—either overall or in specific services—change over time, and if so in what ways?

Answering these questions would require several design, data collection, and analysis/reporting steps for the participation analysis, if added to the basic impact evaluation design, and might include some or all of the following:

1. Identify a source of information on the pool of *potential* JSA participants in each study site—likely the TANF caseload rolls as documented in administrative data. The size of the pool, relative to the number of participants in the impact sample in a site, will answer the first question above regarding the extent of JSA participation in the target population.
2. Inventory the background characteristics data available from each source, which—for those consistently available across sites—will serve to distinguish JSA participants from nonparticipants as per the second specific research questions above.
3. Examine JSA provider agencies' MIS data systems to identify what JSA service delivery variables are recorded for individual participants, data which—for variables consistently available across sites—will allow the analysis to address the third of the three specific research questions listed above.
4. Design a baseline intake form for impact study participants that gathers other important background characteristics of workers to complement administrative data variables in addressing the “who uses each major type of service?” question.
5. Determine how participation study findings can be integrated with those of the other included evaluation components—particularly the impact analysis findings (but also process analysis and benefit-cost analysis findings, if any)—to enhance the learning and policy guidance achieved by the evaluation agenda as a whole.
6. Collect the needed data as specified in the design steps, and ensure they cover the full intake period for the impact analysis sample so that the final specific research question, on changes in participation patterns over time, can be addressed.
7. Conduct the planned analyses and report the results.
8. Relate the participation study findings to the impact analysis results and findings from any other evaluation components to identify potential areas for more effective targeting and service delivery management in the future, as well as possible refinements to the intake and service delivery approaches of TANF JSA programs in general.

As to scale, the coverage of a participation analysis—like that of a process analysis—should mirror the scale of the impact analysis it supports. Typically, one tries to collect data on all the individuals who could have participated in the studied interventions during the intake interval for the impact analysis sample. Since the data all come from administrative sources or a staff-administered intake form, the costs to the evaluation of broad coverage of individuals—as opposed to selective coverage of analytic units in the process study—are minimal, though one cannot be sure that MIS data tracking JSA activity

participation are fully comprehensive of the universe of participants or services received in all cases. The major questions regarding coverage arise at the site rather than individual level: Can the needed data sources be accessed in every site? Is the pool of potential JSA participants identified by those sources conceptually consistent across sites? Are the crucial variables needed to describe *who* participates and in what services available in all sites—or at least enough to make their inclusion in the analysis meaningful for the set of sites as a group?

These issues suggest two critical roles for the evaluation contractor on this component, should ACF choose to move ahead with its design and implementation. The contractor must first determine—and then maximize—the coverage of the data obtained, by site, for data of the different types listed above. It must also develop analytic procedures that can cope with—and provide meaningful interpretation of the participation findings in light of—any remaining gaps in data that remain in terms of sample coverage and missing variables at the site level.

### 7.3 Benefit-Cost Analysis

Policymakers should be interested not just in JSA interventions' relative impacts on participating TANF recipients but also on society as a whole. Benefit-cost analysis seeks to give a full accounting of all the consequences of society undertaking one policy course rather than another—both needed inputs and their costs and resulting outputs and benefits. From such assessments, the following questions can be addressed that greatly broaden the meaning of and policy guidance offered by the impact analysis:

- Will government gain, fiscally, from engaging TANF recipients in JSA services using one of the tested interventions relative to another or relative to the existing program configuration?
- At what level of government will fiscal benefits—or costs—accrue: local, state, or federal? How large are the budgetary consequences in each case?
- Do participants and their families benefit more from one version of the intervention than another? By how much?
- What is the return to society as a whole of adopting the more successful intervention over the alternative? To what extent are up-front costs, such as those of service delivery, offset by later benefits such as lessened assistance use and greater employment and economic output?
- What unmeasured or measured but nonmonetized benefits or costs should be considered? How important would these have to be to turn a favorable monetary “bottom line” for one intervention over the other, from a social perspective, into an unfavorable one?

To answer these questions, a benefit-cost analysis would involve creating and completing a “social accounting matrix.” This matrix, illustrated in Exhibit 7.1, has rows for the benefit and costs elements to be considered and columns for each of the perspectives posed in the research questions above: participants, government at various levels, and society as a whole.

It is essential that comprehensive thinking take place in designing a benefit-cost analysis to ensure that, to the extent possible, all consequences to all social segments are identified and included in the rows of such a matrix—a step we have not attempted to undertake here (offering an initial “partial sweep” of the possibilities). Filling in the cells of the grid is the hardest part of a benefit-cost study, whether filled in (as is preferred where possible) with dollar amounts or with impacts in natural units where not possible (e.g.,

when monetizing benefits is a particular challenge, as in mental health wellness or self-esteem). This requires several design, data collection, and analysis steps:

1. Identifying sources of data that can provide measures of each of the items in the rows of the matrix. In many, but not all, cases the impact analysis will provide this information.<sup>24</sup>
2. Expand the coverage of any participant followup surveys or administrative data sources in the impact evaluation that will also need to contribute outcome measures (not already included in the impact analysis) to the benefit-cost analysis in order to complete the cells.
3. Develop analytic strategies for extrapolating individual benefit and cost items with a larger potential duration than the evaluation's followup period to their long-run equivalents.
4. Establish a social discount rate to bring future benefits and costs into "present value" dollars.
5. Decide on the manner with which to include "society as a whole" in the benefit-cost matrix.
6. Decide on procedures for sensitivity analyses of benefit-cost results to various assumptions made in the course of these calculations, such as the assumed discount rate or the assumed decay rate for projecting long-run benefits.
7. Construct methodologies for doing the calculations necessary to populate the matrix's cells with numbers.
8. Collect the additional supplementary data needed, including JSA program cost data from agency accounting records.
9. Perform the benefit-cost analyses and sensitivity checks and report the results.
10. Relate the benefit-cost findings to the impact analysis results and findings from any other evaluation components to show policymakers the social (and government fiscal) "return" to the government's JSA investment and how that return depends on key categories of effects in the benefit-cost accounting.

---

<sup>24</sup> All of the benefit-cost entries are in fact impact measures, defined conceptually as the difference between outcomes achieved by and resources expended for JSA intervention approach A compared to JSA intervention approach B.

**Exhibit 7.1: Benefits and Costs of JSA Intervention A, Compared to JSA Intervention B, by Accounting Perspective (Table Shell)**

| Benefit or Cost Component  | Participants | Federal Government | State Government | Local Government | Society as a Whole |
|--|--------------|--------------------|------------------|------------------|--------------------|
| Pretax earnings  | \$ ____      | \$ ____            | \$ ____          | \$ ____          | \$ ____            |
| Fringe benefits from work  |              |                    |                  |                  |                    |
| TANF payments  |              |                    |                  |                  |                    |
| TANF administrative costs  |              |                    |                  |                  |                    |
| SNAP (food stamps) benefits  |              |                    |                  |                  |                    |
| SNAP administrative costs  |              |                    |                  |                  |                    |
| Health care costs, including Medicare/Medicaid payments            |              |                    |                  |                  |                    |
| Medicare and Medicaid administrative costs                         |              |                    |                  |                  |                    |
| Payroll taxes  |              |                    |                  |                  |                    |
| Income and sales taxes   |              |                    |                  |                  |                    |
| UI benefits  |              |                    |                  |                  |                    |
| UI administrative costs  |              |                    |                  |                  |                    |
| JSA intervention administrative costs                              |              |                    |                  |                  |                    |
| Work-related expenses (e.g., child care, transportation, clothing) |              |                    |                  |                  |                    |
| Health status  |              |                    |                  |                  |                    |
| Well-being, including leisure                                      |              |                    |                  |                  |                    |
|  |              |                    |                  |                  |                    |
| Labor market effects on third parties                              |              |                    |                  |                  |                    |
| <b>Net Benefits (+) / Costs (-)</b>                                | <b>Σ \$</b>  | <b>Σ \$</b>        | <b>Σ \$</b>      | <b>Σ \$</b>      | <b>Σ \$</b>        |

In terms of scale, by its nature benefit-cost analysis has a comprehensive scope: it seeks to capture *all* the consequences of alternative policy choices wherever they occur in society for all affected units.

Fortunately, this does not mean that all parts of society where consequences may occur have to contribute data to the research. Rather, the scope of the benefit-cost analysis can be limited to the impact evaluation samples of JSA participants and various “samples of convenience” used by secondary studies that provide information needed for impact valuation in dollars (e.g., to translate earnings impacts into impact on taxes paid). Cost data on JSA program operations are the new area of comprehensive data collection that arises from the addition of a benefit-cost analysis component. Collection of this information from agency accounting records is typically attempted for all evaluation sites, though can pose its specific challenges, especially in the case of cost sharing that exists at one-stop centers.

#### 7.4 Recommendations for Expanding the Design and Implementation Effort

From the preceding discussion, it seems apparent that all three potential additions to a DOSE JSA impact evaluation—process, participation, and benefit-cost analysis—would have value to ACF and government policymakers in general. It is our recommendation that if an impact study is implemented in a way that surmounts the challenges and reaches the scale laid out in previous chapters—and if ACF judges such an evaluation to be worth its cost—that ACF undertake a “full-spectrum” evaluation encompassing all four components and the cross-cutting lessons they can impart.

The team that leads any further design work around alternative components for the evaluation should focus particularly on ways in which additional policy issues might be addressed from a combined set of study components.

Given the extent of the likely return to a full-spectrum, highly integrated evaluation the *added* costs of design, data collection, and analysis/reporting for the extra components, relative to the basic investment in implementing the impact component, would in our judgment be:<sup>25</sup>

- Very small for the participation analysis,
- Small for the benefit-cost analysis, and
- Potentially large for the process analysis, depending on the site coverage and data collection model.

Moreover, we do not foresee at this point any obvious technical obstacles to being able to design versions of each of these components that are implementable if a decision is made to move forward. So while it was very prudent for ACF to commission a design and feasibility assessment of the core impact evaluation before funding its implementation, we do not see serious risks in combining design with implementation in undertaking the additional process, participation, and benefit-cost analysis components.

---

<sup>25</sup> Of course, actual levels of effort for each component would have to be estimated by the government in deciding whether to undertake any of these components.

## 8. Conclusion

This is the second document for a design contract from DHHS/ACF/OPRE to Abt Associates on Job Search Assistance (JSA) programs. The first document described current JSA programs and summarized the empirical literature on the impact of those programs (Klerman et al., 2012). This document has discussed options for a new impact evaluation of JSA programs. In particular, ACF seeks a study that would:

- Estimate the impact of JSA program elements, including some combination of the mechanisms through which they operate, the service delivery modalities, and/or specific program components.
- Estimate the relative impacts of multiple variants of JSA programs as they currently exist in the field.
- Estimate effects on a broad range of outcomes, including time to employment, earnings, and welfare benefits, but also other intermediate outcomes and broader measures of household well-being.
- Explore how impacts vary with local context, with two aspects being especially relevant: policy context and economic context.
  - Because TANF programs vary widely—in particular, benefit levels and sanction policy rules—it seems likely that the impact of JSA program designs vary with TANF program context.
  - The low-income labor market is currently extremely weak, but it will recover. Understanding how impacts vary with local economic conditions in both weaker and stronger conditions is important for future program design and targeting.
- Help programs to understand variation in program impacts across selected subpopulations, in order to support decisions about who gets served with which program design and (perhaps) at all.
- Potentially explore the sensitivity of impacts on treated individuals to general equilibrium effects.
- Examine process, benefit-cost, and other complementary analyses to better understand any effects found.

These are valid design goals. Each bullet point corresponds to information that could be used by state TANF programs and JSA programs more broadly.

In addition to these design goals, our design effort proceeded under the assumption that no added resources will be available for reconfiguring JSA programs in the field in conjunction with an evaluation. That said, we assume that some support would be available for sites selected, to encourage their participation in an evaluation. This seems a fine point: on the one hand, no resources will support new programs, but support might be available for evaluation-related activities to induce, or at least assist, sites to be part of an evaluation.

The rest of this concluding chapter revisits the guiding principles for the design process, which we established in the introductory chapter, and discusses each in light of the design options discussed throughout the report. It then addresses the concern of overriding importance emerging from this design appraisal: the sheer size of the impact evaluation needed in order to produce information of policy value.

In closing, an alternative way of achieving this goal at considerably lower cost is offered for ACF's consideration.

## 8.1 The Guiding Principles Revisited: Assessing the Tradeoffs Across Varied Design Options

Four areas of emphasis were identified for appraising the strengths and weaknesses of different impact evaluation designs: research merit, technical merit, practical considerations, and policy relevance. We consider each of these factors in hindsight here, asking how the different design options discussed through the course of the report fare in each domain.

### 8.1.1 Research Merit

As noted in the Introduction, research merit pertains to the evaluation designs' ability to answer the right question(s) for the right population. We consider the tradeoffs across designs along these two criteria.

**Answers the Right Research Question.** All of the designs focus on the causal impact of variation in JSA program design. While an individual-randomized two-arm design can compare two program variants, a multi-arm design obviously can consider more program variants. While a "flat" multi-arm randomized design can consider multiple program variants, a randomized factorial design would also estimate the relative and synergistic effects of the two program variants together. Group randomized designs can, theoretically, answer the same set of research questions, though the unit of analysis differs from designs that randomize individual TANF recipients. Some selected program variants—specifically those with anticipated spillover effects—would be more appropriate for group-randomized designs, whereas most of the JSA program components of interest can be effectively offered through a lottery among eligible individuals and evaluated that way. This is to say, in brief, that all of the design options discussed focus on the right question and thereby would fill a gap in existing knowledge were JSA services to be evaluated accordingly.

**Addresses the Right Population.** This factor considers the evaluation design's ability to achieve external validity, by which we mean produce findings that are a reasonable guide for policy decisions that would affect a characteristic population of TANF recipients. Chapter 4 noted that any of the individual- or group-randomized designs we suggest as relevant would necessarily involve multiple sites, given observations about sample size needs reported in Chapter 6. Among the options for enlisting sites that have distinctive external validity implications are: random selection for national representativeness, casting a wide net and randomizing at the site level, and purposive selection to deliberately consider contextual factors. Without concurrently taking into consideration any practical or technical criteria, on research merit alone the preferred approach would be to use purposive or—if the population of potential sites has adequate dispersion across the key dimensions for which balance and representativeness is thought essential, stratified random—selection of sites to achieve diversity and balance of the research sample compared to a typical state TANF population. However, in as much as an evaluation struggles to recruit the minimum number of sites—as is often true, and may be true here—representativeness may be the enemy of feasibility (though both the Head Start Impact Study and the WIA Gold Standard evaluation included a high proportion of sites, approaching national representativeness).

### 8.1.2 Technical Merit

We now turn to the criteria that comprise technical merit—the extent to which selected designs and related options support the accurate causal attribution of JSA impacts—which includes unbiased

comparisons among JSA programs of interest, adequate sample sizes for statistical reliability, reliable data collection, and data analytic methods.

***Unbiased Comparisons.*** All of the designs examined in this report use random assignment to eliminate the possibility of selection bias and other confounding influences on the impact estimates produced. That is, they have “internal validity.” As is well known, this is the best means of ensuring that an impact evaluation’s findings are unbiased—i.e., that those findings are not skewed up or down in statistical expectation compared to the true direction and magnitude of impact.

***Adequate Sample Size.*** Because of ACF’s interest in the relative effects of variation in JSA program elements—be they specific program components, the modalities through which those components are provided, or the policy mechanisms that collections of components serve—our analysis considers relatively small differential effect size magnitudes, those that might be reasonably expected to occur given the potential policy contrasts under study. We have argued that a plausible minimum detectable differential effect is 0.02 standard deviations. This corresponds to about \$112 in annual earnings. A differential impact any larger from this type of low-intensity intervention seems unlikely. That said, a larger impact may derive from a comparison of interventions that have a starker contrast than two components of common JSA programs. To the extent that greater contrast can be created in the field, smaller sample sizes may support detecting the larger effect that might be generated.

This target MDES (i.e., 0.02) implies an evaluation design that randomly assigns about 50,000 individuals in a two-arm design or about 75,000 in a three-arm design. This is the sample size to estimate the impact of one or two variants of JSA program design, in a single policy or economic context (or averaging over multiple contexts), for the entire population. Estimates of how impacts vary with other JSA program designs, with local economic or policy context, or with individual participant characteristics would require larger samples. Smaller sample sizes increase the magnitude of the smallest effect that can confidently be detected. Detecting an effect of 0.02 standard deviations with a group-randomized design requires more than 1,000 sites and is therefore clearly not feasible. To detect an effect of 0.05 standard deviations—as one might expect for a treatment-control (rather than treatment-treatment) contrast or the kind of high-intensity/low-intensity contrast suggested later in this chapter—would still require more than 300 sites. For this reason, we recommend that attention be focused exclusively on designs with individual random assignment, where the right collection of sites would be able to provide the adequate sample needed to detect differential effects that are small in magnitude, as would be expected of this sort of test of variants in a relatively light-touch intervention.

***Reliable Data Collection.*** Chapter 3 discussed relevant issues of the measurement of constructs and the timing and mode of data collection, with the central point being the following: while a full-sample survey might be optimal for reliably measuring some variables of interest (including in-program participation, time to initial employment, wages, hours, and job type/quality), a summary measure of earnings summarizes much of the labor market experience—hourly wage, hours when working, periods working—in units reflecting the economic importance of the results to the individuals involved and to society. Earnings is also an outcome measure that can be reliably collected from administrative data sources at much lower cost than other labor market measures (which require followup interviews with participants). Operating under a budget constraint, ACF will need to prioritize which will be more important: the power that comes with full coverage (administrative data or an expensive survey) or the possible added nuance that comes with more refined measures (a subsample survey, for example).

Another data collection-related tradeoff pertains to the period over which impact of the JSA interventions will be calculated and the temporal “granularity” of the available outcome measures. These issues have major implications for the timing and mode of data collection. Administrative data can provide quarterly intervals, but that might not be fine-grained enough to capture the experience of six weeks of job search activity, followed by six weeks of added job search, followed by starting a new job, all of which could take place in a single calendar quarter. For that information, a survey is necessary, though we would encourage creative thinking about the type of survey needed (for example, a simple weekly text—“Are you working? Reply Y or N”—would capture continuous weekly information and could be less expensive than a one-time phone survey and more reliable as well, as even a six-month survey would involve some recall error).<sup>26</sup> If it is important for cost or other reasons to confine the evaluation sample to a single followup survey, it could take place early and capture in-program experiences of JSA participants and their very short-term outcomes and impacts from a three-, four-, or six-month followup interview, for example. Alternatively, the focus could be put on longer-term labor market impacts after perhaps a year. In consideration of these tradeoffs, we would urge use of administrative data for longer-term employment and earnings measures, turning instead to a survey to collect data on those shorter-term horizon measures.

A final point about data collection pertains to baseline measures. The individual randomized evaluation designs may require individual consent to participate in a study, at which point baseline data is collected on the full sample. This would provide an opportunity to collect rich information about study participants that would inform future analyses, in part of subgroups but also of dosage effects and/or participants’ experiences of variation in job search assistance program components and/or modalities as predicted by baseline characteristics.

**Data Analytic Methods.** Chapter 4 detailed the analyses that accompany estimating policy impacts across several design options where, in brief, the treatment-control (really treatment-treatment) indicator captures the contrast’s impact. Additional baseline variables increase the precision of those impact estimates, permitting slightly smaller sample sizes than would be the case were no covariates measured. While there may be some specific considerations—e.g., varying the functional form to account for outcome measurement (continuous, binary, categorical, time-varying)—the basic approach is uncontested. We see no tradeoffs across the design options that should concern ACF: data from each can be reasonably analyzed to estimate a causal treatment effect.

### 8.1.3 Practical Considerations

The next category of criteria for assessing this report’s design options is that of practical considerations. Even if designs have research and technical merit, varied practical issues warrant consideration to

---

<sup>26</sup> It is not the charge of this document to create a complete data collection plan, but we highlight here the possibility for thinking unconventionally about data collection methods that might be appropriate for an evaluation of variation in JSA programs. With standard telephone-with-in-field-followup surveys being as expensive as they are, and a text costing 5 cents, it would behoove a future contractor involved in this research to be able to at least consider, if not implement, innovative data collection options. Brief elaboration on this point: should this approach prove feasible, one could consider a weekly text of a small number (two to four) rotating questions, such as: (1) “Are you working? Reply Y or N”; (2) “How many hours did you work last week? Reply with number, or zero if not working” and so on. As noted in Chapter 6, additional pilot testing on the feasibility of this otherwise untested approach would be advisable.

determine the extent to which it would be plausible to carry out the study. To address this practicality criterion, we elaborate on program cost neutrality, site selection issues, and evaluation-related costs.

**Program Cost Neutrality.** We recognize that, very likely, no additional support for program development will be available. This is in line with ACF’s interest in studying JSA programs as they exist in the world, except that in order to know whether something *works* it needs to be positioned in contrast to something that it is not. Our designs have suggested a comparative program contrast, where one or more variants of JSA are evaluated with reference to a different version of JSA. This kind of evaluation design actually has the strong potential to be cost-neutral from the program perspective. Sites will likely not choose to participate if they have to add to their programs or serve more individuals without the resources needed to do so, but if sites are willing (see next topic: “Site Issues”), they can reconfigure *access* to the elements of their JSA program in order to create an evaluation contrast of relevance. That is, if a site offers four elements of what it considers to be its JSA program, and if one (or two, for that matter) of these can be accessed through randomly assigning individuals, then that site can provide a useful test of that program element(s) relative to the rest of its JSA program offerings. Simply doing this, sites would spend less money under this scenario, but we would suggest maintaining cost neutrality: using any savings that result from serving fewer individuals within some program component to increase the contrast that would exist between the basic program elements and that selected for evaluation scrutiny, for example by making services for the other group more intensive. The restriction that cost neutrality imposes on no new program development applies unless costs of what is new are offset by what is removed. In the larger perspective, this would not seem to be a problem, since innovation in new program areas is not part of what ACF would want to test in the anticipated evaluation.

**Site Issues.** Two dimensions of “feasibility” regarding likely site selection exist: that sufficient numbers of suitable sites actually exist; and that a sufficient number of cooperative and capable sites exist. A suitable site is one that could be modified for an evaluation in order to create a program-related contrast, channeling access to two or more groups of randomly divided program eligibles. By “cooperative” and “capable” we mean that sites would be administratively able to manage not only the programmatic reconfiguration, as necessary, but also the demands of participating in an experimental evaluation. We elaborate more on some site considerations in the discussion below, drawing there on observations about these other criteria for judging the feasibility of various design options.

**Evaluation Cost.** The costs of an evaluation hinge on sample size (of individuals and/or sites), including randomization of research units, and the data needed to support analysis of the questions of interest. As sketched in Chapter 3, an approach to dealing with the data cost driver is to make the tradeoff to accept measures that come from less expensive (administrative) sources. A survey is not needed for earnings and benefits, which can reliably be drawn from administrative data. We acknowledge that that strategy means abandoning attempts to measure impacts on other outcomes, in particular, time-to-employment. A subsample survey might provide a more affordable way to collect this and other key measures.

Another unavoidable cost is that of setting up randomization (of individuals or sites). With site-level randomization, one randomizes sites, not participants. As a result, only minimal design and oversight of randomization is needed. However, site-level randomization imposes a sizable recruitment penalty. The number of participating sites must be substantially larger than would be needed for a design where multiple sites provide large numbers of individuals to be randomized. If 50 sites could provide a sample of 50,000–75,000 to detect an MDES of 0.02, we might expect a site-level sample of about 1,000 sites to detect a comparable effect.

#### 8.1.4 Policy Relevance

Throughout this document, we have referred to a primary motivator for this research: usefulness to program design and policy decisions. The policy-relevant contrast is the one that should be the focus of any future evaluation of JSA. To the extent that government agencies—including federal and state administrations as well as county/local agencies—are interested in learning about the relative effectiveness of specific components and/or modes of JSA service delivery, the specific interventions to be tested must be configured accordingly. As some of our discussion has highlighted, a more fundamental question might be of interest: does doing intensive JSA make enough of a difference over doing the bare minimum JSA to justify its costs? ACF will need to set policy priorities, in light of research considerations, to address this question and the overall issue of policy focus if it is to proceed with a JSA impact evaluation based on the current document.

Based on this analysis of design options across criteria, it seems apparent that if we were to suggest following all of the *ideal* evaluation design parameters listed at the beginning of this chapter, the resulting study would be infeasible. Large sample size implies multiple sites, and these demands drive technical considerations, with data collection demands driving cost: a comparative design powered to detect a small differential effect demands large numbers of individuals, which thereby requires multiple sites. Recruiting enough sites to enroll sufficient numbers of cases will be challenging, especially given the minimal inducements that the evaluation contractor could offer to potential sites and the extremely limited influence of ACF over state TANF programs.

Moreover, in order to test a single JSA program component, sites would need to be operating a roughly comparable base program, carving out a roughly comparable component to be tested. This is unlike NEWWS or ERA. In those programs, except for the Labor Force Attachment versus Human Capital Development head-to-head test, the alternative was “no treatment,” so to be comparable, each site only needed to adopt one common program. Here, each site would need to have two common elements, a common standard program and a comparable and separable component. In addition, in those evaluations, a stand-alone analysis was done for each site; that is, single-site samples were sufficient to detect the relevant treatment-control contrast.<sup>27</sup> In this case, we need similar programs to generate sufficient sample size across sites to conduct the main analysis comparing multiple variants of an intervention. Any single site could not be large enough to detect a differential effect on the order of our expectations.

Assuming that sufficient numbers of suitable, cooperative, and capable sites are recruited, costs concerns remain. To survey tens of thousands of people using standard survey methods would cost many millions of dollars. Setting up and overseeing randomization at the number of sites required to generate the needed sample sizes would probably have similar costs. Together, this implies a very expensive evaluation, probably as expensive as anything that ACF has ever fielded (e.g., in current dollars, several times as expensive as NEWWS). Unless a greater contrast (and therefore large MDES) between treatment options could be created, sample sizes will be large.

How should ACF proceed? One approach is direct. ACF might decide that the goal of identifying the impact of varying some specific component or components of JSA programs for TANF recipients is sufficiently important as to warrant the expense. Alternatively, ACF could consider a different approach

---

<sup>27</sup> Then, much later as a followup step, the data across sites were pooled for a meta-analysis.

that would be feasible at a much lower cost. Costs are driven by the large samples and the demands these place on evaluation setup and monitoring costs in hundreds of sites and the large number of individuals who must be interviewed to obtain outcome data from a followup survey or surveys. The need for large samples is driven in turn by the small *difference in impacts* expected between two JSA interventions that incorporate different program elements. Required sample sizes would drop sharply if the expected impact were larger; for example, slightly more than tripling the expected differential impact would cut the required sample size to a tenth of its former value. This suggests a two-pronged strategy for reducing study costs substantially.

***Choose a Sharp Contrast:*** To this point, ACF has deliberately left open the specific aspects of the JSA interventions it may wish to evaluate. That being the case, the experimental contrast in the evaluation could be between high resource intensity and low resource intensity, rather than between specific intervention models. This perspective suggests inducing large variation in the resources allocated to the JSA effort for different subsets of TANF recipients separated at random. By design (if the two groups are of equal size, which we recommend), testing very little against almost twice the usual resource investment would be cost neutral from the JSA program cost perspective. The details of what to do with the resources, both on the high and low ends, would be left up to the site/state. Thus, any site could adopt these design alternatives. However, the evaluation would be asking sites to implement two programs corresponding to the high- and low-intensity options. In the current budget environment, it is hard to imagine sites/states agreeing to do this without some funds for program development—in particular, funded staff time for program managers and supervisors to develop rules for implementing the high- and low-intensity versions of their current JSA approach.

The point of this design would be to determine *if the extent of job search support provided to TANF recipients makes a difference* when pushed to the limit: a great deal versus very little. If not, it is hard to imagine that different configurations of JSA components would distinguish themselves from one another in a less sharply contrasted head-to-head test. And either way ACF will have examined the first principle of policy assistance to disadvantaged families—do something that matters—and it will have done so in an affordable manner. Our greatest concern given all the design exploration in this report is that an expensive evaluation not be undertaken where little (or no) differentiation of results seems likely. The closer the tested interventions are to a “black/white” contrast, the greater the chances that society will learn that how JSA is approached, from a depth and intensity standpoint, makes a difference. That said, future thinking must consider the ethics of reducing what is available to some individuals in the effort to test whether a greater intensity of services makes a difference.

***Reconsider Survey Data Collection:*** Much, but not all of, the incremental cost of a larger sample goes to survey data collection. For JSA, quarterly employment and earnings and monthly TANF benefits are recorded in administrative data that are available at essentially zero marginal cost. If ACF were willing to focus on those outcomes that are available in administrative data, then evaluation costs would drop sharply. It would be feasible to field a smaller survey that would not support labor market impact estimates but could focus on determining the service differentials between the intervention groups in order to characterize the contrast in treatments underlying those impacts (as measured from UI wage data).

Even if ACF accepted a much smaller survey, many incremental costs of evaluation size would remain. Many sites would need to be recruited. In each of those sites, randomization would need to be designed, taught, and monitored. Getting the required sample size down remains important.

## 8.2 Conclusion

We strongly agree with ACF's original insight that JSA services need to be subjected to rigorous impact evaluation in the TANF domain. For most TANF recipients, JSA is the first work-focused activity undertaken; it is also a core component of many other government and private programs. Improving our understanding of the differential impact of different JSA approaches has the potential to improve TANF programs and thereby to improve the lives of TANF entrants and simultaneously to lower TANF benefit costs and TANF caseloads.

However, JSA is a comparatively light-touch program, and existing variation in TANF JSA programs appears to be small. Detecting the small likely differential impacts of such small variation in TANF JSA programs requires very large samples, leading to likely infeasible site recruitment requirements and likely infeasible costs. A major reason for this is the small expected *difference* in how much two different JSA models can boost earnings. If the realistic expectation, as assumed here, is that this difference will be on the order of \$100 per family per year, is that size of economic improvement really worth discovering? Is it worth mounting a large and very expensive evaluation to investigate?

Should ACF conclude that estimating such small impacts is not worthwhile, a study contrasting interventions with substantially larger potential for differential impacts could be undertaken—in particular, the high-intensity versus low-intensity contrast discussed in the previous section. In as much as policy-relevant impacts need to be larger and, through a sharper intervention contrast become more likely, sample sizes could be much smaller. If sample sizes come down, so do costs. As importantly, the required number of sites and required length of sample enrollment also come down. Combined with a decision to survey only a small subset of those randomized, this strategy would yield what appears to us to be a cost-feasible and worthwhile evaluation.

## References

- Abt Associates. (2012). *Revised data collection and analysis plan: Family Options Study*. Bethesda, MD.
- Angrist, J.D. (2006). Instrumental variables methods in experimental criminological research: What, why, and how. *Journal of Experimental Criminology*, 2, 23–44.
- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91, 444-472.
- Banerjee, A., and Duflo, E. (2008). *The experimental approach to development economics* (CEPR Discussion Paper 7037). London: Centre for Economic Policy Research.
- Bell, S. H., Olsen, R. B., Orr, L. L., and Stuart, E.A. (2011). *Estimates of bias when impact evaluations select sites purposively*. Presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington D.C.
- Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8, 225–246.
- Bloom, H. S., Hill, C. J., and Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, 22, 551–575.
- Bos, J., Crosby, D., Duncan, G.J., Gibson, C., Granger, R., Huston, A.C., McLoyd, V., Mistry, R., Magnuson, K., Romich, J., and Ventura, A. (2001). Work-based antipoverty programs for parents can enhance the school performance and social behavior of children. *Child Development*, 72, 318-336.
- Calmfors, L. (1994). Active labor market policy and unemployment—A framework for the analysis of crucial design features. *OECD Economic Studies*, 22, 7–47.
- Caplan, R. D., Vinokur, A. D., Price, R. H., and van Ryn, M. (1989). Job seeking, reemployment, and mental health: A randomized field experiment in coping with job loss. *Journal of Applied Psychology*, 74, 759–769.
- Collins, L. M., Murphy, S. A., Nair, V. N., and Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30, 65–73.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2012). *Do labor market policies have displacement effects? Evidence from a clustered randomized experiment* (NBER Working Paper 18597). Cambridge, MA: National Bureau of Economic Research.
- Davidson, C., and Woodbury, S. (1993). The displacement effect of reemployment bonus programs. *Journal of Labor Economics*, 11, 575–605.
- Ferracci, M., Jolivet, G., and van den Berg, G. J. (2010). *Treatment evaluation in the case of interactions with markets* (IZA Discussion Paper 4700). Bonn: Institute for the Study of Labor.
- Fishbein, M., and Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

- Gautier, P., Muller, P., van der Klaauw, B., Rosholm, M., and Svarer, M. (2012). *Estimating equilibrium effects of job search assistance* (IZA Discussion Paper No. 6748). Bonn: Institute for the Study of Labor.
- Greenberg, David, Robert H. Meyer, Charles Michalopoulos and Michael Wiseman. (2003). Explaining Variation in the Effects of Welfare-to-Work Programs. *Evaluation Review*, 27(4) 359–394.
- Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., Adams-Ciardullo, D., and Gassman-Pines, A. (2001). *National evaluation of welfare-to-work strategies: How effective are different welfare-to-work approaches? Five-year adult and child impacts for eleven programs*. New York: MDRC.
- Heckman, J. J., Hohmann, N., Smith, J., and Khoo, M. (2000). Substitution and dropout bias in social experiments: A Study of an influential social experiment. *Quarterly Journal of Economics*, 115, 651–694.
- Hsieh, C., and Urquiola M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile’s voucher program. *Journal of Public Economics*, 90, 1477–1503.
- Kalil, A. and Leininger L.J. (2008). Cognitive and non-cognitive predictors of success in adult education programs: Evidence from experimental data with low-income welfare recipients. *Journal of Policy Analysis and Management*, 27, 521-535.
- Klerman, J., Juras, R., and Nisar, H. (2012). Estimation of intra-class correlation in the analysis of job training programs. Unpublished draft manuscript. Bethesda, MD: Abt Associates.
- Klerman, J., Koralek, R., Miller, A., and Wen, K. (2012). *Job search assistance programs: A review of the literature*. Cambridge, MA: Abt Associates.
- Lise, J., Seitz, S., and Smith, J. (2004). *Equilibrium policy experiments and the evaluation of social programs* (NBER Working Paper 10283). Cambridge, MA: National Bureau of Economic Research.
- Meyer, B. (1995). Lessons from the U.S. Unemployment Insurance experiments. *Journal of Economic Literature*, 33, 91–131.
- Michalopolous, C., and Schwartz, C. (2000). *What works best for whom: Impacts of 20 welfare-to-work programs by subgroup*. Washington, DC: U.S. Department of Health and Human Services, Office of Assistant Secretary for Planning and Evaluation and Administration for Children and Families; and U.S. Department of Education.
- Moulton, B.R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32, 385-397.
- Moulton, B.R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 72, 334-338.
- Navarro, D., Azurdia, G., and Hamilton, G. (2008). *A comparison of two job club strategies: The effects of enhanced versus traditional job clubs in Los Angeles*. New York: MDRC.
- Nisar, H., Juras, R., and Klerman, J. (2012). *Power parameters for job training programs*. Paper presented at the Welfare Research and Evaluation Conference, Washington, DC.

- Olsen, R., Bell, S., and Luallen, J. (2007). *A novel design for improving external validity in random assignment experiments*. Paper presented to the Association for Public Policy Analysis and Management, Washington, DC.
- Olsen, R. B., Orr, L. L., Bell, S. H., and Stuart, E. A. (2012). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, 32, 107–121.
- Parisi, D., McLaughlin, D. K., Grice, S. M., and Taquino, M. T. (2006). Exiting TANF: Individual and local factors and their differential influence across racial groups. *Social Science Quarterly*, 87, 76–90. DOI: 10.1111/j.0038-4941.2006.00369.x
- Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, 24(2), 157–187.
- Peck, L. R., and Mayo, A. (2011). *An empirical assessment of external validity*. Presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC.
- Pissarides, C. (2000). *Equilibrium unemployment theory*, 2<sup>nd</sup> ed. Cambridge, MA: MIT Press.
- Price, R. H., van Ryn, M., and Vinokur, A. D. (1992). Impact of a preventive job search intervention on the likelihood of depression among the unemployed. *Journal of Health and Social Behavior*, 33, 158–167.
- Price, R. H., and Vinokur, A. D. (1995). The Michigan Jobs Program: Supporting career transitions in a time of organizational downsizing. In M. London (Ed.), *Employee development and job creation: Human resource strategies for organizational growth*. New York: Jossey-Bass.
- Raudenbush, S. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173–185.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D.B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, 2(1), 1-26.
- Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review*, 33, 537–567.
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.
- Tipton, E., and Hedges, L. V. (2011). *Sample selection in randomized experiments: A new method using propensity score stratified sampling*. Presented at the Annual Fall Research Conference of the Association for Public Policy Analysis and Management, Washington, DC.

- Vinokur, A. D., Price, R. H., Caplan, R. D., van Ryn, M., and Curran, J. (1995). The Jobs I preventive intervention for unemployed individuals: Short- and long-term effects on reemployment and mental health. In L.R. Murphy, J.J. Hurrell, Jr., S. L. Sauter, & G. P. Kota (Eds.), *Job stress interventions* (pp. 125-138). Washington, DC: American Psychological Association.
- Vinokur, A., and Caplan, R. D. (1987). Attitudes and social support: determinants of job-seeking behavior and well-being among the unemployed. *Journal of Applied Social Psychology*, 17, 1007–1024.
- Wanberg, C. R. (1997). Antecedents and outcomes of coping behaviors among unemployed and reemployed individuals. *Journal of Applied Psychology*, 82, 731–744.
- Wanberg, C. R., Kanfer, R., and Rotundo, M. (1995). Unemployed individuals: Motives, job-search competencies, and job-search constraints as predictors of job seeking and reemployment. *Journal of Applied Psychology*, 84, 897–910.
- Wanberg, C. R., Zhang, Z., and Diehn, E. W. (2010). Development of the “Getting Ready for Your Next Job” inventory for unemployed individuals. *Personnel Psychology*, 63, 439–478.

## Appendix A: Bias in Experimental Impact Measures when JSA Participants Displace Other Workers: Schematic Analysis and Possible Solutions within the Experiment

Begin by assuming that a local economy contains two types of low-skilled workers who might be provided with JSA through a TANF-focused program: Type A workers who ideally fit the available low-skill jobs if they can be assisted in finding them and Type B workers who fit those jobs less well and hence have lower productivity and earnings when working at those jobs. For simplicity, assume all jobs of relevance to Type A and Type B workers are the same and that all Type A workers would earn  $Y$  dollars per week on those jobs while all Type B workers would earn  $X$  dollars per week on those jobs, with  $X$  smaller than  $Y$  ( $X < Y$ ). Also for simplicity, presume that no other jobs obtainable by either type of worker—with or without JSA—exist in the economy.

These assumptions imply the schematic of a local labor market (LLM) shown in Exhibit A.1, an LLM that includes 114 jobs of the relevant type (“CXIV” in Roman numerals)<sup>28</sup> and many more low-skilled workers of Types A and B. The exhibit lists all the job slots to be filled down the left margin, from Job I to Job CXIV, and then shows—across its columns—various ways those jobs might be filled by the low-skilled workers. Further rows in the exhibit correspond to workers who are not successful in obtaining any job. Workers of the two types are designated by “A” and “B” in the body of the table, and by “A[#]” and “B[#]” where specific workers need to be tracked through the analysis. Asterisks (\*) indicate the workers who receive JSA in particular scenarios.

The columns of the exhibit show labor market outcomes—who holds which jobs for what pay—under three different scenarios:

- A world with low-intensity JSA services (the “Low-Intensity JSA” columns), which inefficiently allocates many of the available jobs to less productive Type B workers while Type A workers go jobless;
- A world with high-intensity JSA services provided to eight individuals (the “High-Intensity JSA” columns), services that create somewhat more efficient job matches by placing A4\* in a job (Job CXII) that would otherwise have been held by B2; and
- A world in which 16 people participate in a randomized experiment (the “JSA Experiment” and “Experimental Impact” columns) to gain access to the eight high-intensity JSA slots.

Impacts are derived for each of these scenarios in other columns of the exhibit and are discussed below.

### A.1 The Effect of More Effective JSA on Job Holding and Earnings Absent an Experiment

The first six columns of Exhibit A.1 map jobs to workers and earnings in two scenarios: where low-intensity JSA is provided (the “Low-Intensity JSA” columns) and where high-intensity JSA is provided to eight people (the “High-Intensity JSA” columns), four of them Type A and four of them Type B

<sup>28</sup> The points made here generalize to any number of jobs.

(designated by \*). Under either scenario, Jobs I and II at the top of the exhibit are filled with Type A workers who receive JSA when it is available but in fact do not need that degree of assistance to obtain those jobs. Moving down the exhibit to the last two jobs shown, Jobs CXIII and CXIV are filled with Type B workers (B3 and B4) who receive intensive JSA when available but who also do not need that degree of support to obtain those jobs. Two additional Type B workers near the bottom of the exhibit receive intensive JSA when available but end up without jobs just as they would have with low-intensity JSA.

Of greater interest are the two additional Type A workers in the middle of the exhibit (A3 and A4) who receive high-intensity JSA when available but who would not have worked without it. Of these, intense JSA is effective in placing one of them in a job: person A4 holds job CXII and earns  $Y$ , replacing a Type B worker (B2) who would have held that job and earned  $X$ . This generates an earnings impact of  $Y-X$  as the only change in earnings for the LLM economy as a whole. Dividing this amount by the eight workers receiving high-intensity JSA, we get an average differential earnings impact of  $\frac{(Y-X)}{8}$  per high-intensity JSA participant, as shown at the bottom of the sixth column of the exhibit.

## A.2 Impact Estimates from an Experiment with Individual-Level Random Assignment

Now consider running a random assignment experiment to allocate the eight high-intensity JSA slots among individuals. Here, we need to identify 16 people to include in the random assignment pool, since eight will be assigned to the low-intensity JSA slots. The final six columns of Exhibit A.1 identify these cases. The workers who would have filled the high-intensity JSA slots in the absence of the experiment—which we will, following Olsen, Bell, and Luallen (2007), call the “program preferred” for high-intensity group—are underlined here and designated as in either the high-intensity JSA group (H) or the low-intensity JSA group (L). As can be seen, these are the individuals who would have received low-intensity JSA normally in a program that had the capacity to deliver high intensity to only half of those it served and program staff got to choose which TANF recipients to route into which level of services. These are the \* cases in the left half of the exhibit, only half of whom receive high-intensity JSA in the experiment (those designated with underlining and \*) since the other half are picked at random for the low-intensity group.

Another eight workers appear in the “H” and “L” columns of the exhibit. These individuals *would not have received high-intensity JSA absent the experiment*. We call this set of workers the “next tier applicant” (NTA) group to signify their “next best” status in the eyes of program intake staff deciding which JSA applicants to serve with their limited number of high-intensity JSA slots. These workers get the opportunity to participate in high-intensity services when “the doors are opened wider” for that option by the experiment—i.e., when it includes 16 individuals in a 50:50 lottery to fill the eight available high-intensity slots. This group consists of A1, A2, A5, and A6 near the middle of the exhibit plus two other type A individuals near the top and two type B individuals near the bottom. Each of these pairs is divided between the H and L columns and the H member of each pair receives high-intensity JSA services (\*) while the other member does not.<sup>29</sup>

<sup>29</sup> Note that we deliberately construct a pool of added cases that differs from the program preferred pool in its composition: a 6:2 ratio of Type A to Type B workers for the added group, a 4:4 ratio for the program preferred

The final three columns of the exhibit show the earnings impacts captured by the experiment. The experiment combines earnings with positive signs for H group members with earnings with negative signs for L group members to give a net aggregate earnings gain of  $2Y+X$ , as shown near the bottom right of the exhibit. This translates into an average impact on the eight high-intensity JSA participants in the experiment of  $\frac{(2Y+X)}{8}$ . The parts of this impact accruing to the different subpopulations are also shown at the bottom right of the exhibit. For the program preferred group of four high-intensity JSA participants the aggregate impact is  $Y+X$  and the average impact  $\frac{(Y+X)}{4}$ . For the lower-priority group of four high-intensity JSA participants the aggregate impact is  $Y$  and average impact  $\frac{Y}{4}$ . Note that these two separate aggregates sum to the overall aggregate impact of  $2Y+X$  and that the two separate average impacts average to the overall average of  $\frac{(2Y+X)}{8}$ .<sup>30</sup>

### A.3 Displaced-Worker and Extra-Worker Bias in the Standard Experimental Estimate

Crucially, the average impact estimate for the overall experiment,  $\frac{(2Y+X)}{8}$ , does not equal the true average impact per high-intensity JSA participant in the real-world “High-Intensity JSA” scenario shown in the sixth column,  $\frac{(Y-X)}{8}$ . The experiment gives a per-participant impact measure biased upward by  $\frac{(2Y+X)}{8}$ . This bias stems from two sources. To see each of them, first note that the row of the exhibit for Job CXII is the only one that the experiment actually *needs to take account of*, since it is the only place where high-intensity JSA changes outcomes between the “Low-Intensity JSA” and “High-Intensity JSA” real-world scenarios: it puts A4\* in a job that would have been held by B2. This means we need a high-intensity experimental group with a type A person earning  $Y$  and a low-intensity experimental group with a type B person earning  $X$ ; if we had nothing else in the experiment this would give the right aggregate earnings gain. We have the former—A4\* is in the high-intensity group earning  $Y$ . We do not have the latter, since no one in the low-intensity group earns  $X$  (all the Type B individuals in this group are depicted as nonemployed). This is because the earnings loss between the “Low-Intensity JSA” counterfactual and the “High-Intensity JSA” world is experienced by an *out-of-sample individual*, B2, whom the exhibit shows in the “Not in Sample” column of the “JSA Experiment” panel indicating that B2 is in neither of the experimental groups. B2 is displaced from job CXII and the experiment misses the social loss of  $X$  dollars of earnings to that individual. We will call this the “displaced-worker bias.”

But that is not where the bias ends. The high- and low-intensity experimental groups contain other earners beyond those needed to represent the real-world total earnings effect. In some instances, job-holding by these individuals *artificially* differs between the two experimental groups. Examining the rows near the center of Exhibit A.1 under the “JSA Experiment” banner (the rows for jobs CXII through CXIV), we see that the high-intensity group from the experiment—in addition to the  $Y$  dollars of earnings of A4\*—

group. It cannot be assumed that the types of individuals who receive priority for high-intensity services are the same in their composition as those who fall at the next priority level. A different alternative ratio could have been assumed without changing the results of the analysis.

<sup>30</sup> In particular,  $\frac{(Y+X)/4}{2} + \frac{Y/4}{2} = \frac{Y}{8} + \frac{X}{8} + \frac{Y}{8} = \frac{(2Y+X)}{8}$ .

contains  $X$  dollars of earnings for  $B3^*$  and  $Y$  dollars in earnings for  $A1^*$ . The counterparts to these last two workers in the low-intensity group— $B4$  and  $A2$ —do not have jobs, even though there are no true effects on employment and earnings for this set of workers when examined within the “Low-Intensity JSA” and “High-Intensity JSA” columns on the left of the exhibit. Within the experiment, there are no additional earners in the low-intensity group to offset the “unwanted” earnings of the  $B3^*$  and  $A1^*$  workers on the high-intensity side. As a result, an additional upward bias of  $Y+X$  occurs for the high-intensity group and hence for the experimental impact estimate. We call this the “extra-worker bias.”

The sum of the displaced-worker bias of  $X$  and the extra-worker bias of  $Y+X$  results in a total bias of  $Y+2X$  for the experimental impact estimate. This gives an aggregate impact measure from the experiment of  $2Y+X$ , as shown near the bottom right of the exhibit, compared to the desired true total impact of  $Y-X$ , and a commensurate upward bias in the experiment’s average impact estimate.

#### A.4 Correcting the Bias Using a Modified “Program Preferred” Experiment

The distinction between program preferred participants and other participants in the experimental sample provides a means of removing both elements of bias. Olsen, Bell, and Luallen (2007) introduced the program preferred randomized design into the literature and demonstrated its feasibility when randomizing students to the Upward Bound program. A critical feature of the design (abstracted in the discussion here, for simplicity) is the use of a higher T:C ratio for candidates program staff identify as program preferred for better/stronger services than for other candidates. This creates an incentive for intake staff to accurately identify program preferred individuals, since by definition these are the individuals they would most like to serve in the strongest possible way and hence whose odds of “winning the random assignment lottery” they would most like to improve. Note that the only condition that has to be met is that in a world in which high-intensity services are constricted, these are the individuals the program would serve. It does not matter if they would in fact benefit more from the high-intensity services, the direction of displacement, or the nature of selection, e.g., the basis could be provider “hunches” about individuals or formal decision rules. Moreover, this designation of who is program-preferred has to occur *prior* to randomization to affect the lottery odds, making the preferred and nonpreferred cases distinct *exogenous* subgroups of the full experimental sample in the same sense as men and women or high school dropouts and high school graduates are exogenous subgroups. As a result, it is straightforward to estimate impacts on earnings separately for the two groups simply by splitting the random assignment pools (high- and low-intensity cases) symmetrically along those lines.

The final two columns of Exhibit A.1 show how the two resulting impact estimates are constructed and—at the bottom right of the exhibit—give the total impact and average impact estimates for both subpopulations. On average, JSA’s impact on the program preferred subgroup ( $Impact_{PP}$ ) is  $\frac{(Y+X)}{4}$  and its impact on the less preferred subgroup ( $Impact_{NTA}$ ) is  $\frac{Y}{4}$ . These two estimates can be combined in the following linear combination to obtain an unbiased estimate of the average differential impact of the interventions that define the two experimental arms:

$$Impact_{other} - \frac{Impact_{PP}}{2} = \frac{Y}{4} - \frac{Y}{8} - \frac{X}{8} = \frac{Y}{8} - \frac{X}{8} = \frac{(Y-X)}{8}$$

A statistical test of this linear combination of independent, split-sample estimates of  $Impact_{other}$  and  $Impact_{PP}$  is also straightforward to calculate based on the standard error of the linear combination.

## A.5 The Challenge of Unbiased Estimation with Broader Worker Displacement

Exhibit A.1 has one special feature not yet mentioned: it assumes that A1\* benefits from intensive JSA in the experimental design by taking a job (Job CXIV) that would have been occupied by a Type B person *who receives* JSA (person B4\*) in the “High-Intensity JSA” world. It is possible that A1\* will instead displace a Type B worker who would not receive high-intensity JSA in the “High-Intensity JSA” scenario—creating a larger out-of-sample worker displacement problem than previously noted. This situation is shown in the shaded rows of Exhibit A.2. Only three rows in this exhibit differ from Exhibit A.1: the Job CXI row, the Job CXIV row, and the first row without a job. A1\* displaces worker B1 in Job CXI, where B1 is not part of the experimental sample and has no job (i.e., s/he ends up in the first row that contains no job). Job CXIV, which A1\* occupied in the previous exhibit, now continues to be held by B4 just as in the “Low-Intensity JSA” and “High-Intensity JSA” worlds. This results in the same displaced-worker bias of  $Y-X$  as in the previous exhibit (due, as before, to A4\* replacing B2 in Job CXII), but in a smaller extra-worker bias since the extra  $X$  dollars in earnings for B3\* in the high-intensity group are now offset by an extra  $X$  dollars in earnings for B4 in the low-intensity group. Hence, the only extra-worker bias is the extra  $Y$  dollars in earnings for worker A1\*, and the total bias for the experimental impact estimate declines from  $Y+2X$  to  $Y+X$ . This gives a total impact measure from the experiment of  $2Y$ , as shown near the bottom right of Exhibit A.2, compared to the desired true total impact of  $Y-X$ .

Ironically, although the bias is smaller in the case of no overlap in the experimental sample between Type A job takers and Type B job losers, we can no longer correct that bias by using the subgroup strategy. Because Type B workers have exactly offsetting earnings in the high-intensity and low-intensity groups of Exhibit A2 (B3\* earns  $X$  in the high-intensity group, B4 earns  $X$  in the low-intensity group), the experiment can no longer provide information on the magnitude of  $X$ . Indeed, its overall and subgroup average impact estimates all equal  $\frac{Y}{4}$  (see bottom right of the exhibit). Getting from  $\frac{Y}{4}$  to the desired true average impact of  $\frac{(Y-X)}{8}$  without information on  $X$  is impossible. However, dividing the overall average impact estimate of  $\frac{Y}{4}$  by 2 will move the estimate *closer* to the true impact:  $\frac{Y/4}{2} = \frac{Y}{8}$ , which exceeds the true average impact of  $\frac{(Y-X)}{8}$  by  $\frac{X}{8}$  rather than by  $\frac{Y}{8} + \frac{X}{8}$  (as did the original overall average impact estimate of  $\frac{Y}{4}$ ).

Other methodologies for estimating the differential impact of high- versus low-intensity JSA (or any two other JSA intervention models) in the face of job shuffling—when not captured perfectly by standard experimental impact analysis—will be necessary in the non-overlap case. Further work along the lines of this appendix could gain more generalizable results. It will also be important to be able to determine whether the no-overlap condition or the overlap condition applies in a real experiment, a further area for future work.

**Exhibit A.1: Job-Holders, Earnings, and Impacts with Overlapping Displacement in the Experiment**

| Job            | Low-Intensity JSA |      | High-Intensity JSA |      | True Impact | JSA Experiment        |                      |                            | Experimental Impact |                            |                         |
|----------------|-------------------|------|--------------------|------|-------------|-----------------------|----------------------|----------------------------|---------------------|----------------------------|-------------------------|
|                | Person            | \$\$ | Person             | \$\$ |             | High-Intensity Sample | Low-Intensity Sample | Not in Experimental Sample | Total               | Program Preferred Subgroup | Lower-Priority Subgroup |
| I              | A                 | Y    | A*                 | Y    | 0           | A*                    |                      |                            | Y                   | Y                          |                         |
| II             | A                 | Y    | A*                 | Y    | 0           |                       | A                    |                            | -Y                  | -Y                         |                         |
| III            | A                 | Y    | A                  | Y    | 0           | A*                    |                      |                            | Y                   |                            | Y                       |
| IV             | A                 | Y    | A                  | Y    | 0           |                       | A                    |                            | -Y                  |                            | -Y                      |
| V              | A                 | Y    | A                  | Y    | 0           |                       |                      | A                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| LX             | A                 | Y    | A                  | Y    | 0           |                       |                      | A                          |                     |                            |                         |
| LXI            | B                 | X    | B                  | X    | 0           |                       |                      | B                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| CX             | B                 | X    | B                  | X    | 0           |                       |                      | B                          |                     |                            |                         |
| CXI            | B1                | X    | B1                 | X    | 0           |                       |                      | B1                         |                     |                            |                         |
| CXII           | B2                | X    | A4*                | Y    | Y-X         | A4*                   |                      |                            | Y                   | Y                          |                         |
| CXIII          | B3                | X    | B3*                | X    | 0           | B3*                   |                      |                            | X                   | X                          |                         |
| CXIV           | B4                | X    | B4*                | X    | 0           | A1*                   |                      |                            | Y                   |                            | Y                       |
| --             | A1                | 0    | A1                 | 0    | 0           |                       | B4                   |                            | 0                   | 0                          |                         |
| --             | A2                | 0    | A2                 | 0    | 0           |                       | A2                   |                            | 0                   |                            | 0                       |
| --             | A3                | 0    | A3*                | 0    | 0           |                       | A3                   |                            | 0                   | 0                          |                         |
| --             | A4                | 0    | B2                 | 0    | 0           |                       |                      | B2                         |                     |                            |                         |
| --             | A5                | 0    | A5                 | 0    | 0           | A5*                   |                      |                            | 0                   |                            | 0                       |
| --             | A6                | 0    | A6                 | 0    | 0           |                       | A6                   |                            | 0                   |                            | 0                       |
| --             | A                 | 0    | A                  | 0    | 0           |                       |                      | A                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| --             | A                 | 0    | A                  | 0    | 0           |                       |                      | A                          |                     |                            |                         |
| --             | B                 | 0    | B*                 | 0    | 0           | B*                    |                      |                            | 0                   | 0                          |                         |
| --             | B                 | 0    | B*                 | 0    | 0           |                       | B                    |                            | 0                   | 0                          |                         |
| --             | B                 | 0    | B                  | 0    | 0           | B*                    |                      |                            | 0                   |                            | 0                       |
| --             | B                 | 0    | B                  | 0    | 0           |                       | B                    |                            | 0                   |                            | 0                       |
| --             | B                 | 0    | B                  | 0    | 0           |                       |                      | B                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| --             | B                 | 0    | B                  | 0    | 0           |                       |                      | B                          |                     |                            |                         |
| Total impact   |                   |      |                    |      | Y-X         |                       |                      |                            | 2Y+X                | Y+X                        | Y                       |
| Average impact |                   |      |                    |      | (Y-X)/8     |                       |                      |                            | (2Y+X)/8            | (Y+X)/4                    | Y/4                     |

KEY: \* = Worker receives high-intensity JSA; \_\_ = Worker is in the program preferred subgroup of intervention candidates.

Exhibit A.2: Job-Holders, Earnings, and Impacts with No Overlapping Displacement in the Experiment

| Job            | Low-Intensity JSA |      | High-Intensity JSA |      | True Impact | JSA Experiment        |                      |                            | Experimental Impact |                            |                         |
|----------------|-------------------|------|--------------------|------|-------------|-----------------------|----------------------|----------------------------|---------------------|----------------------------|-------------------------|
|                | Person            | \$\$ | Person             | \$\$ |             | High-Intensity Sample | Low-Intensity Sample | Not in Experimental Sample | Total               | Program Preferred Subgroup | Lower-Priority Subgroup |
| I              | A                 | Y    | A*                 | Y    | 0           | <u>A</u> *            |                      |                            | Y                   | Y                          |                         |
| II             | A                 | Y    | A*                 | Y    | 0           |                       | <u>A</u>             |                            | -Y                  | -Y                         |                         |
| III            | A                 | Y    | A                  | Y    | 0           | A*                    |                      |                            | Y                   |                            | Y                       |
| IV             | A                 | Y    | A                  | Y    | 0           |                       | A                    |                            | -Y                  |                            | -Y                      |
| V              | A                 | Y    | A                  | Y    | 0           |                       |                      | A                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| LX             | A                 | Y    | A                  | Y    | 0           |                       |                      | A                          |                     |                            |                         |
| LXI            | B                 | X    | B                  | X    | 0           |                       |                      | B                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| CX             | B                 | X    | B                  | X    | 0           |                       |                      | B                          |                     |                            |                         |
| CXI            | B1                | X    | B1                 | X    | 0           | A1*                   |                      |                            | Y                   |                            | Y                       |
| CXII           | B2                | X    | A4*                | Y    | Y-X         | <u>A4</u> *           |                      |                            | Y                   | Y                          |                         |
| CXIII          | B3                | X    | B3*                | X    | 0           | <u>B3</u> *           |                      |                            | X                   | X                          |                         |
| CXIV           | B4                | X    | B4*                | X    | 0           |                       | <u>B4</u>            |                            | -X                  | -X                         |                         |
| --             | A1                | 0    | A1                 | 0    | 0           |                       |                      | B1                         | 0                   | 0                          |                         |
| --             | A2                | 0    | A2                 | 0    | 0           |                       | A2                   |                            | 0                   |                            | 0                       |
| --             | A3                | 0    | A3*                | 0    | 0           |                       | <u>A3</u>            |                            | 0                   | 0                          |                         |
| --             | A4                | 0    | B2                 | 0    | 0           |                       |                      | B2                         |                     |                            |                         |
| --             | A5                | 0    | A5                 | 0    | 0           | A5*                   |                      |                            | 0                   |                            | 0                       |
| --             | A6                | 0    | A6                 | 0    | 0           |                       | A6                   |                            | 0                   |                            | 0                       |
| --             | A                 | 0    | A                  | 0    | 0           |                       |                      | A                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| --             | A                 | 0    | A                  | 0    | 0           |                       |                      | A                          |                     |                            |                         |
| --             | B                 | 0    | B*                 | 0    | 0           | <u>B</u> *            |                      |                            | 0                   | 0                          |                         |
| --             | B                 | 0    | B*                 | 0    | 0           |                       | B                    |                            | 0                   | 0                          |                         |
| --             | B                 | 0    | B                  | 0    | 0           | B*                    |                      |                            | 0                   |                            | 0                       |
| --             | B                 | 0    | B                  | 0    | 0           |                       | B                    |                            | 0                   |                            | 0                       |
| --             | B                 | 0    | B                  | 0    | 0           |                       |                      | B                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| .              | .                 | .    | .                  | .    | .           |                       |                      | .                          |                     |                            |                         |
| --             | B                 | 0    | B                  | 0    | 0           |                       |                      | B                          |                     |                            |                         |
| Total impact   |                   |      |                    |      | Y-X         |                       |                      |                            | 2Y                  | Y                          | Y                       |
| Average impact |                   |      |                    |      | (Y-X)/8     |                       |                      |                            | Y/4                 | Y/4                        | Y/4                     |

KEY: \* = Worker receives high-intensity JSA; \_ = Worker is in the program preferred subgroup of intervention candidates.