Social Experiments in the Labor Market

Jesse Rothstein[*]
University of California Berkeley

Till von Wachter
University of California Los Angeles

Chapter prepared for the Handbook of Experimental Economics.

PRELIMINARY. PLEASE DO NOT CIRCULATE.

Abstract

Large-scale social experiments were pioneered in labor economics, and have been used to study topics ranging from the effect of job training to incentives for job search to labor supply responses to taxation. Yet, many questions routinely asked in the context of social experiments in labor economics require going beyond random assignment. This includes questions pertaining to both internal and external validity, including endogenously observed outcomes, such as wages and hours; spillover effects; site effects; heterogeneity in treatment effects; multiple and hidden treatments; and the mechanisms producing treatment effects. In this Chapter, we review approaches that address these design issues in the context of randomized control trials in labor. These approaches expand the range of questions that can be answered using experiments by combining experimental variation with econometric or theoretical assumptions. We also discuss efforts to build the means of answering these questions into the ex ante design of experments. Our discussion yields an overview of the expanding toolkit available to experimental researchers.

[*] Contact: rothstein@berkeley.edu, tvwachter@econ.ucla.edu. We thank Ben Smith for sterling research assistance.

I. Introduction

There is a very long history of social experimentation in labor markets, primarily in the United States. These experiments have addressed core labor market topics such as labor supply, job search, and human capital accumulation, and have been central to the academic literature and policy discussion for many decades.

By many accounts, the first large-scale social experiment was the New Jersey Income Maintenance Experiment, initiated in 1968 by the Office of Economic Opportunity to test the effect of income transfers and income tax rates on labor supply. In contrast to recent experimental practice, this experiment was designed to map out a response surface rather than to evaluate a specific program. Participants were assigned to a control group or to one of eight different treatment arms that varied in the income guarantee to a family that did not work and the rate at which this was taxed away as earnings rose. Three follow-up experiments – in rural North Carolina and Iowa; in Gary, Indiana; and in Seattle and Denver, with varying benefit levels and tax rates (and, in Seattle and Denver, a cross-cutting set of counseling and training treatments) – were begun before data collection for the New Jersey experiment was complete.

Other early, labor market experiments examined the effects of job search encouragement for Unemployment Insurance recipients; job training and job search programs; subsidized jobs for the hard-to-employ; and programs designed to push welfare recipients into work (Greenberg and Robins, 1986). These topics have been

returned to repeatedly in the years since, as researchers have sought to test new program designs or to build on the limitations of earlier research. There have also been many smaller-scale experiments, on bonus pay schemes, management structure, and other firm-level policies.[1]

From the beginning, the use of random assignment experiments (also known as random-control trials, or RCTs) has been controversial in labor economics. The primary appeal of RCTs is that they solve the assignment, or selection, problem in program evaluation. In non-experimental, or observational, studies, program participants may differ in observed and unobserved ways from those who do not participate, and econometric adjustments for this selection rely on unverifiable, often implausible assumptions (Lalonde 1986; Fraker and Maynard 1987; though see also Heckman and Hotz, 1989). With a well-executed randomization study, however, the treatment and control groups are comparable by design, making it straightforward to identify the effect of the treatment under study.

But set against this very important advantage are a number of drawbacks to experimentation. Early on, it was recognized that RCTs can be very expensive and hard to implement successfully. Ideally, everyone assigned to receive a treatment should receive a full dose, and those assigned to the control group should be barred from any treatment. But this is often not possible. Sometimes it is not feasible to control participants' behavior, and many participants deviate from their intended treatment assignments. In other cases, ethical, political, or operational

[1] We omit here audit studies aimed at uncovering discrimination in the labor market and elsewhere (e.g., Bertrand and Mullainathan 1994; Kroft, Lange, and Notowidigdo 2013). These are covered by Bertrand and Duflo, elsewhere in this volume.

considerations make it undesirable to limit access to alternative treatments. Although this can be partly addressed within the basic experimental paradigm, it does limit what can be learned.

More importantly, critics of over-reliance on experimentation point to a range of important questions that are not answered via randomization alone. We consider a number of such questions in this chapter. Here is a partial list, with examples of situations in which they arise:

- *Questions about impacts on endogenously observed outcomes*. Consider the effect of job training on wages. Because wages are observed only for those who have jobs, and because training may affect the likelihood of this, even the contrast in mean wages between randomly assigned treatment and control groups does not compare like to like and thus does not solve the assignment problem for this outcome.
- *Questions about spillovers and market-level impacts*. When one individual's outcome depends on others' treatment assignments, experimental estimates of treatment effects can be misleading about a program's overall effect. In the context of labor market programs, an increase in search effort by a treatment group may lower the chances to find jobs of the control group, leading to an overstatement of the program's total effect (which will depend importantly on the scale at which it is implemented). Similar issues can arise if subjects communicate with each other, leading to a dilution in treatment.
- *Questions about heterogeneity of treatment effects*. Experiments have limited ability to identify heterogeneity of treatment effects, especially if

heterogeneity is not fully characterized by well-defined observable characteristics. This is often of first-order importance, as in many cases the relevant question is not *whether* to offer a program (e.g., job training) but *for whom* to make it available, or *which* versions of the program are most effective (and why).

- *Questions about generalizability*. While in ideal cases experiments have high internal validity for the effect of the specific program under study on the specific experimental population, in the setting in which it is studied, they may have limited external validity for generalizations to other locations, to other programs (or even to other implementations of the same program), or to other populations. For example, a reemployment bonus program may have a very different effect in a full-employment local economy than when the local area is in a recession, or the same program offered in different sites may have dramatically different effects due to variation in local program administration or context.
- *Questions about mechanisms*. Many questions of interest in labor market research do not reduce to the effects of specific "treatments" on observed outcomes, but relate, at best, to the mechanisms by which those effects arise. For example, an important question for the analysis of unemployment insurance programs is whether the unemployed are liquidity constrained or whether they can borrow or save to smooth consumption optimally across periods of employment and unemployment. And important questions about the design of welfare and disability policy turn on whether observed non-

employment is due to high disutility of work or to moral hazard. Neither of these questions can be boiled down to the effect of a treatment on an observed outcome. Carefully designed experiments can shed light on the phenomena of interest, but may not be able answer them directly.

To be clear, all of these questions are thorny under any methodological approach, and are generally no easier to answer in quasi-experimental studies than in randomized experiments. Many critics of experimentation point to the importance of identifying the "structural" parameters – a full characterization of program enrollment decisions and the process that determines observed outcomes – that determine program selection and impacts. Many of the design issues above could be avoided or addressed with estimates of the underlying structural parameters. But these structural parameters are difficult to measure. So-called "structural" methods generally trade off internal validity in pursuit of more external validity, but a study that fails to solve the assignment problem is unlikely to yield any more external than internal validity.

Unfortunately, however, it is rarely possible to design an experiment that directly identifies the structural parameters of interest. Thus, there is often value in combining the two paradigms. This involves imposing untestable assumptions about the processes of interest, while still resting on experimentation (or other empirical methods that offer high internal validity) where possible. The additional assumptions can dramatically enhance external validity if they are correct, though if they are incorrect – and this is generally untestable – both internal and external validity suffer.

The current frontier for labor market research thus involves combining the best features of the two approaches to permit answers to more questions than are addressed by simple experiments while retaining at least some of the credibility that these experiments can provide.

In this chapter, we discuss a variety of questions that require this sort of approach. We distinguish two broad strategies for answering these questions using experimental data. First, one can augment traditional experiments by imposing additional structure, either economic or econometric, after the fact. In many cases, the amount of structure required, and the strength of the additional assumptions that are necessary, is small relative to the value of the results that can be obtained. Section IV of this chapter discusses a number of examples where scholars have used this approach fruitfully. This includes analyses of issues such as endogenously observed outcomes (e.g., Ahn and Powell 1993, Grogger 2009, Lee 2009), hidden treatments (e.g., Kline and Walters 2014, Feller, Grindal, Miratrix, and Page 2014, Pinto 2014), heterogenous treatment effects (e.g., Kline and Walters 2014, Heckman and Vytlacil 2005), and multiple treatments and mechanisms (e.g., Card and Hyslop 2005, Schmieder, von Wachter, and Bender 2014, Della Vigna, Lindner, Reizer and Schmieder 2014). Our review gives a snapshot of an expanding toolkit with which researchers can address a wider range of questions based on variation from RCTs.

The second broad strategy is to address the limitations of traditional experiments ex ante, via design of the experimental intervention or evaluation itself. In many cases, clever design choices – multiple treatment arms, carefully designed stratification, or randomization both across and within groups, for example – can

allow for much richer conclusions than would be possible via traditional experiments. This sort of approach has a very long history – indeed, the very first large-scale social experiments, the income maintenance experiments of the late 1960s and early 1970s – can be seen as a version of this strategy. But the pendulum swung away for a long time, and researchers have only recently begun to return to experimental designs that synthesize random experimental variation with structural modeling. Recent examples of this approach include Kling, Liebman, and Katz (2006) who use it to address potential biases from endogenous attrition, and Crepon, Duflo, Gurgand, Rathelot, and Zamora (2013), who quantify the importance of spillovers. In our view, approaches like these represent the current research frontier.

The rest of this chapter proceeds as follows. In Section II, we give a brief overview over the history of social experiments in the labor market, focusing on the type of programs and questions that have been analyzed. Section III summarizes the econometrics of RCTs, and gives an overview of potential remaining design issues. There, we also compare RCTs to other approaches used to answer some of the questions typically addressed by social experiments in the labor market. Section IV then reviews studies imposing econometric or theoretical structure to expand the range of questions that can be answered by existing RCTs. Section V discusses the potential of using the design of experiments to resolve potential design issues ex ante. The remaining sections VI and VII provide a more detailed overview of existing social experiments.

II. A Primer on the History and Topics of Social Experiments in the Labor Market

As the "credibility revolution" has swept over the field of empirical economics in the last generation, the role and status of experimental evidence has grown.

- *NIT is the first experiment*
  - *Not a program evaluation – maps out a response surface, and explicit goal is to understand income and substitution effects.*
  - *Differs from current practice in a number of ways: Assignment probabilities differ across strata, many treatments.*
  - *Limited by attrition.*
  - *Analyses of NIT data nearly all go beyond simple T-C contrasts, to estimate parametric labor supply models (with tobits for nonparticipation and in some cases structural models of nonlinear tax schedules).*
  - *But field changes after NIT, in part in response to perceived shortcomings.*
- *Basic facts about frequency, from Greenberg and Shroder*
- *Common topics studied*
  - *Human capital development*
    - *Job training for disadvantaged workers*
  - *Labor supply*
    - *Welfare-to-work*
    - *Job search incentives for unemployed workers*
    - *Responses to taxation*
  - *Job search and matching*
- *Patterns:*
  - *Evaluations of specific programs*
  - *One (or sometimes two) treatment arms.*
  - *Focus on T-C contrasts*
  - *Explicit standards for attrition*
  - *Often examine endogenously observed outcomes, conditional on observation, but recognize the limitations of this.*

III. Social experiments as a tool for program evaluation

Random assignment solves the selection problem that often plagues non-experimental program evaluations, and makes it possible to generate uniquely credible evidence on the effects of well-defined, successfully implemented programs. In the absence of random assignment, people who participate in a program (those who are "treated") are likely to differ in observed and unobserved ways from those who do not participate, and the effect of this selection can be distinguished from the causal effect of the program only via the imposition of strong, unverifiable assumptions about the selection process. But experiments have limitations as well – while they can have very high internal validity, at closer inspection this is true only for certain types of programs and certain types of outcomes; and even then there can be other challenges, such as difficulties in generalizing from the experimental results to a broader setting.

In this section, we discuss the value of experiments as a means of solving the selection problem, then discuss some limitations of the experimental paradigm for program evaluation and policy analysis. Our discussion draws heavily on Angrist-Imbens-Rubin's (1996) "potential outcomes" framework. Some of the limitations we discuss can be addressed via careful design of the experimental study, while others require augmenting experimental methods with other tools. We take up these topics in Sections V and VI, respectively.

a. A brief refresher on the econometrics of randomized experiments

The benchmark case: Experiments with perfect compliance

The appeal of randomized experiments is that they make transparent the assumptions that permit causal inference and create a direct tie from the implementation of the experiment to the key selection assumption. The simple contrast between those assigned to participate in the program and those excluded identifies the effect of being assigned to participate, subject only to the assumption that the randomization was conducted correctly. Moreover, in many cases this effect is identical to the effect of the program on its participants (known as the "effect of the treatment on the treated," or TOT), which is often the main parameter of interest; in other cases, it is straightforward to convert the effect of assignment to participate (often known as the "intention to treat, or ITT, effect) into an estimate of the program treatment effect for a subpopulation of interest.

To illustrate this, and to set up notation that will be useful later, we use Donald Rubin's potential outcomes framework for causal inference as set forth in Holland (1984). We consider the evaluation of a simple, well-defined program, such as job training or a bonus scheme to encourage rapid return to work after a job displacement, where it is possible to assign individuals separately to participate or to be excluded from participation in the program. For each individual *i*, one can imagine two possible outcomes: One that would obtain if *i* participated in the program, y_{1i}, and one that would obtain if he or she did not participate, y_{0i}.[2]

This notation makes it easy to define the effect of the program on the individual. The program's causal effect on person *i* is simply the difference between

[2] This notation rests on a specific assumption about the mechanisms by which the program operates, known as the "stable unit treatment value assumption," or "SUTVA." We discuss SUTVA at greater length below.

the outcome which would obtain if he/she participated and that which would obtain if she did not, $\tau_i = y_{1i}-y_{0i}$. When $\tau_i > 0$, *i* would have a higher outcome if he/she participated than if he/she did not; when $\tau_i < 0$, the opposite is true.

A fundamental hurdle for causal inference is that τ_i cannot be measured directly. Any individual either does or does not participate in the program, so either y_{1i} or y_{0i} can be observed, but not both. Let D_i be an indicator for participation, with $D_i = 1$ if *i* actually participates in the program and $D_i = 0$ if *i* does not. Then we observe only D_i and $y_i = D_i*y_{i1} + (1-D_i)*y_{i0}$.

The simplest estimator of the program's effect is the contrast between the average outcomes of those who participate and those who do not. This is:

$$E[y_i \mid D_i = 1] - E[y_i \mid D_i = 0] = E[y_{i1} \mid D_i = 1] - E[y_{i0} \mid D_i = 0]$$

$$= E[y_{i0} + (y_{i1} - y_{i0}) \mid D_i = 1] - E[y_{i0} \mid D_i = 0]$$

$$= E[y_{i1} - y_{i0} \mid D_i = 1] + (E[y_{i0} \mid D_i = 1] - E[y_{i0} \mid D_i = 0])$$

$$= E[\tau_i \mid D_i = 1] + (E[y_{i0} \mid D_i = 1] - E[y_{i0} \mid D_i = 0]).$$

Thus, the simple participant-nonparticipant contrast combines two distinct components: The effect of the treatment on the treated, $\tau^{TOT} = E[\tau_i \mid D_i = 1]$, and a selection term, $E[y_{i0} \mid D_i = 1] - E[y_{i0} \mid D_i = 0]$, that captures the difference in outcomes that would have been observed between those who participated in the program and those who did not, had neither group participated (for example, had the program not existed). This second term arises because the process by which people select (or are selected) into program participation may generate differences between participants and non-participants other than their participation statuses. If so, the treatment-control difference cannot be interpreted as an estimate of the effect of the

program. The challenge of distinguishing the two terms of (#) is sometimes known as the "fundamental problem of causal inference."

Studies that attempt to identify the causal parameter τ^{TOT} (or any other summary of the distribution of τ_i) must make assumptions about the selection term (or, alternatively, about the process that governs selection into treatment). In simple participant-nonparticipant contrasts, the necessary identifying assumption is that the selection term in (#) equals zero. More complex observational studies (e.g., with regression controls, propensity score matching, instrumental variables, or other designs) may make different assumptions. But all necessarily rely on identifying assumptions that make it possible to distinguish causal effects from selection. And in each case, the estimated causal effects are only as credible as the corresponding assumption.

Explicit random assignment can make the identifying assumption both extremely simple and highly credible. Consider a case where individuals are recruited into a study sample from some underlying population, then randomly assigned either to participate in the program or not to. This random assignment ensures that D_i is independent of $\{y_{0i}, y_{1i}\}$ conditional on being in the study sample. This implies that $E[y_{i0}|D_i = 1] = E[y_{i0} \mid D_i = 0]$, and the treatment-control contrast equals (in expectation) $\tau^{TOT} = E[\tau_i \mid D_i = 1]$. Thus, randomization of D_i permits identification of the causal effect of the program on its participants, τ^{TOT}. Moreover, randomization also ensures that the participant have the same distribution of τ_i as the non-participants, so the TOT effect equals the average treatment effect (ATE),

$E[\tau_i]$, in the population represented by the study sample. Thus, the average causal effect can not only be identified but extrapolated to the larger population.[3]

In a nutshell, this is the value of randomization in program evaluation. In a simple randomized control trial, the identification assumption that justifies causal inference is identical to the assumption that the randomization was correctly executed – that there are no systematic differences between the treated and control samples that would lead to differences in the expectations of y_{0i} or τ_i between them. Of course, in any finite sample there may be differences in the sample averages of y_{0i} or τ_i. But this variation is captured by the standard error of the experimental estimate. The estimate is unbiased, with measurable uncertainty, so long as the groups are the same in expectation.

Imperfect compliance and the local average treatment effect

One complication that often arises, and that will be central to some of our discussion below, is that it is not always possible to control subjects' program participation. Some subjects who are assigned to receive job training may not show up to their course, while others who are assigned to the control group, and thus not to receive training, may find another way into the program. One approach to formalize this is to introduce an additional variable, Z_i, representing the experimenter's intention for individual i: An individual with $Z_i = 1$ is intended to be served, and one with $Z_i = 0$ is not to be. The experimenter can randomly assign Z_i,

[3] This holds if the entire population of interest is part of the experiment. As we come back to below, if the study sample is not representative of the broader population, the ATE identified will be local to the subpopulation represented by the sample.

but, because program participation is not fully controlled, cannot ensure that D_i is randomly assigned. There may be some (non-randomly selected) individuals who are assigned $Z_i = 1$ but wind up with $D_i = 0$, for example if they fail to arrive for their assigned training course, and others who are assigned $Z_i = 0$ but wind up with $D_i = 1$, for example if they locate an alternative training provider. Although Z_i is independent of potential outcomes, due to random assignment, D_i may not be.

We now need some new notation: Let D_{0i} represent the individual's treatment status if assigned $Z_i = 0$ and D_{1i} represent the treatment status if assigned $Z_i = 1$. In a slight abuse of notation, we continue to define the two potential outcomes y_{0i} and y_{1i} in terms of the actual *treatment* status D_i, not the intended status Z_i. Indeed, we assume that Z_i has no effect whatsoever on outcomes except via its effect on D_i – an individual who is assigned to the control group but receives treatment anyway will have the same outcome that she would have had had she been assigned to the treatment group, and the same for those who do not receive treatment. This is a plausible assumption in some cases – for example, when the alternative job training providers offer courses that are identical to those being offered by the program under study. But in other cases it is not at all reasonable – alternative courses may be quite distinct. We discuss this issue at length below, but for now we maintain the assumption of a single binary treatment status D_i that is observed without error.

The contrast in mean outcomes between those meant to be treated and those meant not to be can be written as the sum of three terms:

$$E[y_i \mid Z_i = 1] - E[y_i \mid Z_i = 0]$$

$$= E[D_{1i}y_{i1} + (1-D_{1i})y_{i0} \mid Z_i = 1]$$

$$- E[D_{0i}y_{i1} + (1-D_{0i})y_{i0} \mid Z_i = 0]$$

$$= E[(D_{1i} - D_{0i})(y_{i1} - y_{i0}) \mid Z_i = 1]$$

$$+ (E[D_{0i}(y_{1i} - y_{0i}) \mid Z_i = 1] - E[D_{0i}(y_{1i} - y_{0i}) \mid Z_i = 0])$$

$$+ (E[y_{0i} \mid Z_i = 1] - E[y_{0i} \mid Z_i = 0]).$$

Here, we have two selection terms: Those with $Z_i = 1$ may differ from those with $Z_i = 0$ both in their potential outcomes if not treated, the final term, and in their treatment effects, which enter into the second term multiplied by D_{i0}. A randomly assigned Z is independent not only of potential outcomes but also of potential treatment statuses, so as before random assignment eliminates the selection terms. The simple treatment-control contrast thus reduces to:

$$E[y_i \mid Z_i = 1] - E[y_i \mid Z_i = 0] = E[(D_{1i} - D_{0i})\tau_i \mid Z_i = 1] = E[(D_{1i} - D_{0i})\tau_i].$$

This is known as the "intention to treat" (ITT) effect. It represents the actual effect of offering access to the program in the setting in which the experiment takes place, and in many cases is the effect of policy interest. In other cases, however, one might want to identify the effect of program participation (as distinct from the offer to participate).

When the actual receipt of treatment, D, is observed, it is possible to identify a summary of the effect of the treatment itself. This requires an additional assumption, known as "monotonicity" or the "no defiers" condition.[4] Specifically, we

[4] Other assumptions might suffice in place of monotonicity. For example, a constant treatment effect would be identified by the procedure here even if monotonicity did not hold, so long as $E[D_{i1}-D_{i0}] \neq 0$.

need to assume that there is no one who would participate in the program if assigned to the control group but not participate if assigned to the treatment group: $\Pr\{D_{1i} = 0 \text{ and } D_{0i} = 1\} = 0$. If it holds, $D_{1i} - D_{0i}$ is either zero or one, and the intention-to-treat effect above can be written as:

$$E[(D_{1i} - D_{0i})\tau_i] = E[\tau_i \mid D_{1i} - D_{0i} = 1] \Pr\{D_{1i} - D_{0i} = 1\}.$$

The second term here is identified as the difference in treatment status between those assigned to the experimental and control groups:

$$E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0] = E[D_{1i} \mid Z_i = 1] - E[D_{0i} \mid Z_i = 0]$$

$$= E[D_{1i}] - E[D_{0i}] = \Pr\{D_{1i} - D_{0i} = 1\}.$$

Thus, the ratio of two experimental treatment-control contrasts -- that for outcomes and that for the realized treatment status -- isolates a summary of the τ_i distribution:

$$E[\tau_i \mid D_{1i} - D_{0i} = 1] = (E[y_i \mid Z_i = 1] - E[y_i \mid Z_i = 0])/(E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]).$$

Importantly, this does *not* equal the average treatment effect, $E[\tau_i]$, except in special cases. What is identified here has been called the *local* average treatment effect (LATE) on the subpopulation of people who would receive the treatment if assigned to the treatment group ($D_{1i}=1$) but would *not* receive treatment if assigned to the control group ($D_{0i}=0$). This subpopulation is known as the "compliers" with the experimental assignment. The complier subgroup may or may not be the population of interest, and the complier average treatment effect may differ from the ATE or even from the TOT. For example, in many settings one would expect that people who will receive the largest benefits from treatment to make disproportionate efforts to obtain it, even if assigned to the control group; in this case, the complier average treatment effect will be smaller than the TOT.

Unfortunately, randomization of Z is not sufficient, without further structure, to identify the ATE or TOT.

b. Examples

Many social experiments in the labor market have indeed been of the standard structure as just described.

- *Brief descriptions of:*
  - *JTPA*
    - *Assignment to three "groups" (classroom training, OJT/job search, other) before randomization.*
    - *Control group couldn't receive JTPA services (but could receive others)*
    - *Sites selected non-randomly, with widely varying services.*
    - *Outcomes of interest: Employment, earnings, welfare receipt, educational attainment.*
  - *Illinois UI incentive experiment*
    - *Treatment was eligibility for an employee or an employer bonus; primary outcome was total benefit payment.*
    - *Heterogeneity across sites, also across individuals in eligibility for extended benefits.*
  - *GAIN welfare experiment in California*
    - *Treatment was education, job search, skills training, work experience; population was AFDC recipients.*
    - *Outcomes: Employment-related activity participation; earnings; welfare receipt; employment.*
    - *Heterogeneous impacts across sites (big effects in Riverside).*
  - *Connecticut Jobs First*
    - *Treatment was time limit, enhanced earnings disregard, job search requirement, stronger sanctions.*
    - *Outcomes: Employment; earnings; benefit receipt; child well being.*
- *Limitations of T-C contrasts – questions we'd like to ask*
  - *How well was JTPA targeted?*
  - *Which JTPA sites did well? How well would other sites do?*
  - *Why did GAIN do so much better in Riverside?*

- *What is mechanism of reemployment bonus effect (particularly given low take-up of bonuses)?*
- *Impacts of each program on wages.*
- *Extensive/intensive margin effects of Jobs First.*

c. Limitations of the experimental paradigm

The basic experimental paradigm is invaluable for its ability to resolve the fundamental problem of causal inference, by ensuring that estimated program effects are not confounded by selection into treatment. But it cannot solve all identification problems faced by program evaluators, nor answer all questions posed by labor economists seeking to understand the workings of the labor market.

Consider, for example, the above reemployment bonus program. One might want to know the effect of the reemployment incentive on the quality of the job that a participant finds (as measured by the wage). Because wages are observed only for the non-representative subset of study participants who find jobs, and because the program may affect which participants wind up in this subset, this effect is not identified. Other questions that one might want to ask are whether the program effect would differ if it were expanded to more sites, or if a larger share of the job-seeker population in an area were given the incentive. These sorts of extrapolations are not supported by the simple experimental paradigm.

Other typical treatments of social experiments in labor economics are more difficult to cast within the basic framework. For example, in many experimental analyses of welfare programs, the treatment varies several components of the program – time limits, job search requirements, training activities, child care –

simultaneously. In principle, one could call the composite set of components that treatment-group individuals are offered the "treatment" D_i, and the incidence of working (as opposed to not working) the outcome. However, it would be more difficult to assess which aspect of the program actually causes any employment effect, or whether any such effect reflects labor supply responses or increases in human capital. This greatly complicates the extrapolation of the program effect to other populations, since different individuals may in effect face different incentives and hence different potential outcomes.

Another limitation comes from training program evaluations. In many of these, it is impossible to prevent control group subjects from obtaining training from some other, non-program source. This could be modeled as noncompliance, leading to an instrumental variables estimate of the LATE for subjects who would not obtain training in the absence of the program under study. But the alternative training that is obtained by the control group may be an imperfect substitute for what is provided by the program under study, such that summarizing participation by a single binary D variable is inappropriate.

In the remainder of this section, we will briefly discuss these and related complications of the standard experimental design.

The Stable Unit Treatment Value Assumption

The above brief overview of the econometrics of experiments glosses over an important assumption, known as the "stable unit treatment value assumption," or SUTVA. Intuitively, this assumption states that the outcome of individual i is unaffected by the treatment status of each of the other study participants. For many

program evaluations, this is innocuous. But in other cases, particularly when a program may have general equilibrium effects, it can be quite restrictive. For example, the provision of job search assistance to some individuals may create "congestion" in the labor market, reducing the job-finding rates of others participating in that market. This is a violation of SUTVA, and will lead a simple randomized trial to overstate the total effect of job search assistance. Another potential violation of SUTVA occurs if the treatment group interacts with each other or with the control group in a way that changes the nature or dilutes the treatment.

Endogenously observed outcomes

In many labor market experiments, some outcomes of interest are observed only for a subset of individuals. For example, weekly hours of work (labor supply) and hourly wages are observed only for those who are able to find jobs, not for those who are unemployed. Even ideal experiments with perfect compliance may not identify causal effects of programs on these outcomes.

To illustrate, consider a program aimed at unemployed workers that includes skill development and job search assistance modules. We are interested in whether the program raises the probability that a participant is employed one year after participation and whether it makes them more productive when employed. For simplicity, we assume that participation is randomly assigned and compliance is perfect.

We have two outcomes here. We denote employment status by $y_i = D_i y_{1i} + (1-D_i)y_{0i}$. For those who are employed at the follow-up survey, we observe the wage $w_i = D_i w_{1i} + (1-D_i)w_{0i}$. Treatment effects of the program on the two outcomes are τ^{y_i}

and τ^w_i. (We imagine that w_{di} is well defined for an individual with $y_{di} = 0$, d={0,1}, but simply not observed. In this case w_{di} can be thought of as the individual's *latent* productivity, that which he/she would be paid if a job were found.)

Estimation of $E[\tau^y_i]$ is straightforward, as discussed above. But the impact on wages is much harder. In general, it is not possible to identify the average treatment effect $E[\tau^w_i]$; the treatment-on-the-treated effect $E[\tau^w_i \mid D_i = 1]$; or even the average treatment effect for the subpopulation that would have been employed with or without the program (for whom τ^w_i is least problematic), $E[\tau^w_i \mid y_{0i} = y_{1i} = 1]$.

The problem here is that it is impossible to distinguish, within each D_i group, between those workers who would also have worked in the counterfactual and those who would not have. Consider the treatment-control difference in mean observed wages:

$$
\begin{aligned}
&E[w_i \mid y_{1i} = 1, D_i = 1] - E[w_i \mid y_{0i} = 1, D_i = 0] = \\
&\quad = E[w_{0i} + \tau^w_i \mid y_{1i} = 1, D_i = 1] - E[w_{0i} \mid y_{0i} = 1, D_i = 0] \\
&\quad = E[\tau^w_i \mid y_{1i} = 1, D_i = 1] + (E[w_{0i} \mid y_{1i} = 1, D_i = 1] - E[w_{0i} \mid y_{0i} = 1, D_i = 0]) \\
&\quad = E[\tau^w_i \mid y_{1i} = 1, D_i = 1] + \\
&\qquad + (E[w_{0i} \mid y_{0i} = 1, y_{1i} = 1, D_i = 1] - E[w_{0i} \mid y_{0i} = 1, y_{1i} = 1, D_i = 0]) \\
&\qquad + (E[w_{0i} \mid y_{1i} = 1, D_i = 1] - E[w_{0i} \mid y_{0i} = 1, y_{1i} = 1, D_i = 1]) \\
&\qquad - (E[w_{0i} \mid y_{0i} = 1, D_i = 0] - E[w_{0i} \mid y_{0i} = 1, y_{1i} = 1, D_i = 0]).
\end{aligned}
$$

The first term here is the average treatment effect in the subpopulation that works under treatment. It may not equal the overall average treatment effect, but insofar as the potential wages of those who do not work are not relevant to social welfare, it is arguably the parameter of interest. The second term solely reflects selection into

treatment, and is zero under random assignment. But the third and fourth terms have to do with selection into employment, not selection into treatment. Random assignment does not ensure that they are zero, and the treatment-control contrast among workers may therefore be badly biased.

A quantitative example helps to illustrate the problem. Suppose that there are two types of workers, those who will earn high wages w^H if employed and those that will earn low wages w^L. Suppose we conduct a randomized experimental evaluation of an integrated job training and job search assistance program. 60% of workers will be low productivity with or without the program ($w_{1i} = w_{0i} = w^L$), 20% will be high productivity with or without the program ($w_{1i} = w_{0i} = w^H$), and 20% will be low productivity without the program but will become high productivity if exposed to the training sequence ($w_{0i} = w^L$, $w_{1i} = w^H$). All of the second and third groups will find jobs, with or without search assistance ($y_{0i} = y_{1i} = 1$), but those in the first group of low-skill, impossible-to-train workers will find work if and only if they receive search assistance ($y_{0i} = 0$, $y_{1i} = 1$).

In this setting, the program's average treatment effect on employment is 0.6; the average effect on latent productivity is $0.2*(w^H - w^L)$; and the average effect on wages of those who would work with or without the program is $0.5*(w^H - w^L)$. We would like to design analyses of the experimental data that can recover these values.

The first, the ATE for employment, is simple: This is identified by the simple treatment-control contrast in employment rates when program participation is randomly assigned. But what about the corresponding contrast for wages, restricted to the subsample for whom wages are observed? In the control group, wages are

observed for only the 40% who find jobs, and average $0.5*w^L + 0.5*w^H$. In the treatment group, everyone finds jobs, and the mean wage is $0.6*w^L + 0.4*w^H$. Thus, the estimated treatment effect is $-0.1(w^H - w^L) < 0$. Selection has led to a perverse estimate here: The training program has a positive effect on 20% of participants and a negative effect for no one, but the experiment appears to indicate that it reduces earnings. The problem, of course, is that the treatment-control comparison of wages must condition on employment, and the subpopulation that is employed differs between the two groups. The problem could be resolved if it were possible to identify the sub-group that responds to the job search component of the program and exclude them from the wage analysis, but nothing in the experiment allows us to identify this group. (Note that it is not resolved by including the non-employed in the wage analysis, with wages set to zero. This yields an estimated effect of $0.2w^H + 0.4w^L$ – the right sign, but a very misleading magnitude.) Without an ability to measure *counterfactual* employment status at the individual level, the program effect on wages is not identified.

This situation arises frequently in analyses of labor market programs. Many of the outcomes of interest – not just wages, but also hours of work, career advancement, or retention on the job – are observed only for those who find jobs in the first place. The problem is widely recognized, and estimated program effects on these endogenously selected outcomes are treated only as suggestive. But this substantially limits our ability to conduct complete evaluations of many programs of interest.

External validity (sites, partial variation in T, imperfect compliance, subgroups)

Another large class of limitations in experiments has to do with generalizing beyond the experimental sample. As noted earlier, while experimental analyses can identify the average treatment effect of the program studied in the population represented by the experimental sample (or, with noncompliance, in the complier subpopulation of that population), extrapolations to other programs, other samples, or other treatment regimes can be hazardous.

We will discuss in this paper two broad classes of external validity issues. One derives from differences between the population of interest and that included in the experimental sample – one might want to understand a program's effect on a population that differs from that represented in the experimental sample, or on a subpopulation other than the experimental compliers.

The second broad class has to do with variations in the treatment on offer. In many programs, the treatment is not homogeneous across locations; in other cases, the treatment may be homogeneous but the counterfactual outcome distribution varies across locations. In either case, one might expect the treatment effect to vary, and there are often policy decisions that turn on being able to identify this cross-site variation.

Mechanisms (multiple treatments, hidden treatments, generalizing to other programs)

Several types of external validity concerns

- *JTPA: Lots of potential treatments (training, search assistance, etc.), and the main question is who should be assigned to which? Gets at heterogeneity and also at distinguishing the effects of the different*

components of the omnibus effect of being assigned to be eligible for services.

- *SSP/Jobs First: Also lots of components of the treatment. But here we aren't (primarily) interested in understanding the effect of giving people different combinations of components, but rather understanding the mechanisms by which the effects of the omnibus treatment package arise, for welfare calculations and extrapolation to other settings.*
- *Site effects. Want them for two reasons: To understand how treatment effect varies with context, and as a way of measuring "dosage" due to differences in program implementation (e.g., for purposes of accountability of contracted providers). A closely related point to the first: Experiments are usually carried out at a non-randomly chosen set of sites, and want to generalize to other sites (see Hotz paper in Manski-Garfinkel volume)*
- *Hidden treatments as a type of mechanism – if TE is zero because of hidden treatments (a type of noncompliance) that leads to little contrast in treatment, that has very different implications than if they are zero because neither treatment works.*

d. Quasi-experimental and Structural Research Designs

Quasi experiments can be valuable alternatives to experiments, where the latter are not feasible or too expensive. But they don't do much to help with the above

design issues. Structural methodologies can resolve the design issues, but at substantial cost to internal validity.

IV. Going Beyond Treatment-Control Comparisons to Resolve Additional Design Issues

Whether one is interested in structural parameters or program evaluation, many questions of policy or intrinsic interest in labor and public economics require going beyond the basic RCT design described in Section III.b. We discussed a number of these questions in Section III.d. Here, we discuss a number of studies that extend the basic RCT design, generally by imposing additional structure beyond the minimal assumptions necessary to support an RCT, to provide answers to these questions.

In many of these examples the additional structure imposed is justified by appeal to theoretical considerations and is just sufficient to extend the RCT to address a specific question and the design issue it raises. In that sense, the studies can be viewed as an effort to bridge pure experimental (or quasi-experimental) approaches credibly identifying a limited number of (potentially composite) causal parameters, with more traditional structural estimation that obtains a fuller characterization of the economic problem via the imposition of substantial additional assumptions. In the ideal case, they maintain the best of both worlds. Yet, while they depend on additional untestable assumptions, by construction most studies do not solve more than one of the design issues discussed.

In this Section, we focus on cases where an existing experiment or quasi-

experiment is used to answer a new question. Section V will consider the case in which more complex experiments can be designed, through appropriate ex ante choices, to address a specific question of interest.

We organize our discussion around the major potential design issues we mentioned in Section III.d. For each, we discuss proposed solutions and, where relevant, point out potential extensions and limitations. We begin by discussing studies that address aspects relating to internal validity, including SUTVA violations (e.g., potential general equilibrium effects) and endogenously observed outcomes. We then discuss studies that address external validity concerns, including site and sub-group effects; effects on subpopulations other than experimental compliers; hidden or multiple treatments; mechanisms for treatment effects; and studies of optimal or simply alternative policies. The discussion is mean to highlight the different approaches, as well as to clarify the scope, potential, and difficulties that arise when extending inference from standard RTCs to a broader range of questions.

a. Endogenously observed outcomes

Many salient outcomes of labor market experiments are only observed if an individual is employed. As discussed in Section III.c, the classic example of a program with potentially endogenous outcomes is training programs. A full evaluation of training programs requires an effect of training on wages. But since training may also affect the decision to work, differential selection into employment by the treatment and control groups makes it impossible to obtain unbiased estimates of the effect of training on wages using experimental variation alone. The

longer-term success of welfare-to-work programs is also linked to wage effects. If the program raises worker productivity and hence wages, as predicted by human capital theory, this may be a key channel through which the benefit of working – and hence labor force attachment – remains elevated after program-based work incentives expire. Similarly, we are often concerned whether welfare programs – and tax changes more generally – affect hours worked (or taxable earnings) of those employed. But again, analyzing changes in labor supply at the intensive margin is hampered by selection bias if the extensive margin varies as well.

There are other examples of salient outcomes of experiments that can only be observed conditional on individual choices, and are hence potentially endogenous. For example, program benefits that are based on family size, assets, or labor force status may affect fertility, savings, or employment, respectively, and hence who participates in the program. Another example is that collection of data from the treatment and control group may be differentially affected by non-response or attrition, especially if the treatment has longer-term effects.

Non-random attrition in particular has been a long-standing concern in the experimental literature in labor economics (e.g., Hausman and Wise 1979). A classic experimental design would be deemed successful if attrition is low, and if present, it is balanced in terms of magnitude and observable characteristics between the treatment and control groups. If this is the case, reweighting the samples may still recover the effect of the TOT or LATE among the original set of compliers (e.g., Ham and Li 2011). Yet, there are relatively few explicit attempts in the literature to address selection bias in other contexts. In contrast, a large literature in labor

economics has dealt with sample selection problems, especially in the analysis of wages and hours in the context of the classic human capital and labor supply models. Largely based on that literature, here we will review several approaches to deal with selection bias: the use of control functions to address selection; estimation of percentiles effects instead of mean impacts; use of additional data to control for selection; construction of bounds based on selection probabilities; construction of bounds using theory.

Parametric selection corrections

The 'classic' approach to control for selection bias in estimating the effects of treatment effects on wages or hours worked is based on control functions. Labor supply theory is used to derive an expression for the selection bias in the outcome equation. This bias depends crucially on the participation equation, which determines the degree of sample selection (e.g., Gronau 1974). Under the assumption of joint normality of the errors in the outcome and selection equations, one can obtain an explicit functional expression of the bias term in the outcome equation. Under these assumptions, unbiased estimates can be obtained either by jointly estimating the full system by maximum likelihood, or by proceeding in a stepwise fashion (e.g., Heckman 1979).

Early on it was recognized that unless experimental variation in participation (e.g., an exogenous instrument affecting only participation and not the outcome equation) is available, identification is only based on functional form assumptions, and results can be quite misleading if these assumptions are even slightly incorrect. By contrast, a substantial literature has shown that once an instrument for

participation is available, treatment effects in the outcome equation can be identified under quite general functional form and distributional assumptions (e.g., Newey, Powell, and Walker 1990). For example, Ahn and Powell (1993) show that under assumptions of a single, strictly monotonic index for selection, variation in the probability of participation independent from the variables in the outcome equation suffices to control for selection.[5] The difficulty is, of course, that often such independent source of variation is not available.[6] We will come back to this in Section V, when we discuss examples how the experimental design itself may be modified to obtain exogenous variation in participation.

Non- and semi-parametric selection corrections

Absent such an instrument, in the presence of selection the treatment effect on the mean outcome is not identified. However, several studies have exploited the fact that under certain assumptions quantile-treatment effects (QTEs), such as the median, may be consistently estimated even in presence of selection. A QTE for the q-th quantile is defined as the difference in the q-th quantile of the outcome distribution in the treatment and control groups, respectively.[7] It is not necessary to

---

[5] In an experimental context, Card and Hyslop (2005) and Manoli and Looney (2014) compare employment rates in the treatment and control groups to assess the potential for selection when studying the effect of the Canadian Self-Sufficient Project (see Section IV.g) and the EITC on wages, respectively.

[6] Card and Hyslop (2005) show that under two additional assumptions, the effect of a work subsidy on wages on wages is still identified even in the presence of selective employment. They show that if the program only has positive effects on labor supply and does not affect the wages for those that would have worked anyways, then the experimental effect on the hourly wage can be consistently estimated by the ratio of the treatment effect on total earnings divided by the treatment on total hours worked.

[7] For any random variable Y having cumulative density function $F(y) = Pr[Y< y]$, the qth quantile of F is defined as the smallest value, such that $F(y_q) = q$. If we consider two distributions F_0 and F_1, then $QTE(q) = y_q(1)- y_q(0)$, where $y_q(g)$ is the quantile of distribution F_g.

observe each individual's outcome to compute the q-th quantile; it suffices to know that someone is above or below that quantile. Thus, if one can assume that all those who are not employed have potential wages in the bottom q percent of the distribution, one can estimate the treatment effect on the qth quantile of potential wages by merely assigning all non-workers the minimum observed value (e.g., Powell 1984, Buchinsky 1994).

It is not clear, however, that the required assumption holds – at any given time, some high-wage individuals may be nonemployed (e.g., Altonji and Blank 1999). Moreover, this strategy is only useful in so far as differences in quantiles of the outcome are deemed sufficient for evaluating the effect of the program.

Johnson, Kitamura, and Neal (2000) note that in longitudinal data, the minimum of all observed wages for a given individual provides an upper bound for the reservation wage. Their approach points to the value of longitudinal data for resolving this selection problem, as outcomes in periods when they are observed are informative about potential outcomes that are not.

Another approach uses reservation wages to measure selection into the subsample of observed wages. This works because – if correctly measured – the reservation wage captures the lowest wage for which an individual is willing to work. Hence, the reservation wage provides the censoring point for an individual's wage offer distribution, allowing one to make inferences about potential wages for those individuals not working in the treatment and control group. Grogger (2009) uses reservation wage information from a randomized evaluation of Florida's Family Transition Program, a welfare-to-work program with emphasis on work

incentives and time limits. With this information, he estimates the treatment effect of the program on wages using a bivariate, censored regression model that allows for classical measurement error in both observed wages and reservation wages. Once Grogger (2009) controls for selection, he finds the program had statistically significantly positive effects on wages.

Addressing the selection problem using direct measures of reservation wages makes intuitive use of the reservation wage concept. Moreover, often information on reservation wages is already being collected in the context of programs providing job search assistance, or if not they are at least in principle relatively easy to elicit if the experimental design includes a survey component. However, recent research suggests that in practice reported reservation wages appear to only partly reflect the properties of the theoretical concept (e.g. Krueger and Mueller 2014), casting some doubt on the robustness of this approach. In particular, Krueger and Mueller report that a substantial number of workers accept (reject) jobs offering wages below (above) their reservation wage, implying that care should be taken in using reservation wages of the nonemployed to make inferences about unobserved wage offers.

Yet another approach is to attempt to bound the selection bias under minimal assumptions, not attempting to obtain a point estimate but rather to investigate how severe the bias from selection could possibly be, under conditions more general than the monotonicity assumption inherent in the Ahn and Powell (1993) and similar estimators.

One bounding approach is proposed by Horowitz and Manski (2009). This strategy asks how much the estimated treatment effect would be inflated if all missing treatment observations were assumed to have the highest possible outcomes and all missing control observations the lowest, then how much it would be depressed if the opposite assumptions were made. Unfortunately, these bounds are typically not very tight, particularly when the outcome variable's support is potentially unbounded. Lee (2009) proposes a strategy for obtaining tighter bounds, via stronger assumptions: He assumes that anyone not employed in the control group would also have been non-employed had they been in the treatment group, so that selection bias arises solely from participants it the treatment group who are employed but would not have been had they been assigned to control.[8] He can then bound the treatment effect by making extreme assumptions about this latter group. Denote the excess fraction employed in treatment group by p. The upper (lower) bound is constructed by removing the lowest (highest) fraction p observations from the treated subsample and recomputing the mean outcome for the treatment group – effectively making the worse case assumption that selection was fully responsible for the entire upper or lower tail of values. Lee (2009) shows that the resulting bounds are sharp and provides formulas for the standard errors. In the case of Job Corps, the procedure results in informative bounds suggesting positive wage effects from training – albeit a zero effect is contained in the confidence interval.

The Lee (2009) approach requires relatively weak assumptions. It presumes only that selection is monotonic in the treatment – that treatment only increases, or

[8] The role of treatment and control groups are reversed if the treatment reduces employment.

only reduces, selection into employment. Monotonicity is implied by standard empirical binary choice models typically used to model participation choices (e.g., Vytlacil 2002), and hence bounds based on trimming are applicable to a wide range of problems, including selective employment, survey non-responses, or sample attrition.

If one is willing to impose further structure from theory, one may obtain bounds more specific to a particular problem. This is especially useful if the theory has explicit predictions about how the endogenous outcome responds to incentives.[9] This is pursued by Kline and Tartari (2014), who analyze the randomized evaluation of Connecticut's Jobs First welfare-to-work program. While previous analyses had found only small responses in hours (the intensive margin), absent an instrument for participation (the extensive margin) sample selection makes such estimates hard to interpret. Kline and Tartari (2014) use revealed preference arguments in the context of a canonic but non-parametric static labor supply model to describe which responses at the intensive and extensive margin to the treatment are consistent with the theory. A facilitating aspect is that due to the nature of the program, absent further structure the treatment can only distinguish between a limited set of discrete counterfactual employment transitions. The result is a mapping of discrete counterfactual outcomes under treatment and non-treatment. The question then is how likely certain transitions are, and in particular

[9] This may be more easily done for hours, which is typically assumed to be a choice variable, than for wages. Yet, to some degree wage may be a choice variable as well, for example if jobs offer wage and effort combinations among which workers choose. This is the approach taken in modern public finance, which often substitutes taxable earnings as choice variable for labor supply.

whether changes at the intensive and extensive margin occur with positive probabilities. Since Kline and Tartari (2014) can only use the marginal distribution among the counterfactual states for the treatment and control groups to infer the transition probabilities between counterfactual states, they cannot point-identify the transition probabilities. Instead, they construct bounds for the transition probabilities among the entire (discretized) distribution of counterfactual states, including the probability of changes in the intensive margin due to the treatment. Their approach also allows them to test the restrictions from the model.

This approach is useful, since it allows Kline and Tartari (2014) to infer about intensive margin responses to the Jobs First program in the presence of selection. Their results could also be used to think about the likelihood of intensive margin responses for similar programs in similar populations. Alternatively, using the marginal distribution of the existing program (AFDC, the program of the control group), the estimated bounds for the matrix of transition probabilities could be used to construct bounds for the intensive and extensive labor supply responses that could arise if Jobs First was implemented at another site. A potential issue is that the procedure is somewhat complex and specific to Jobs First, and hence more representative of a general approach rather than a method applicable to a range of problems. Nevertheless, since many social experiments are concerned with welfare and other programs that may provide explicit variation in incentives and hence useful information on the likelihood of counterfactual outcomes, it is useful to

consider the role that theory can play in providing bounds on treatment effects on endogenous outcomes.[10]

b. Spillover Effects and other Failures of SUTVA

Social experiments in labor economics typically occur in the context of the local or regional labor market. This means that if the number of workers participating in the program is sufficiently large with respect to the relevant segment of the labor market, the program could have an effect on the labor market outcome of the treatment group. As discussed in Section III.c, this would lead to a failure of SUTVA and a potential bias in estimated treatment effect.[11] While this is unlikely to be a problem for the cases in which the treatment group is relatively small with respect to the relevant labor market, the potential extent of such spillover effects would nevertheless matter if the treatment being evaluated is a pilot for a potentially larger-scale program, since any spillover effect would have to be included in a welfare assessment of the program.

Despite its potential prevalence in social experiments in the labor market, relatively few studies have directly dealt with the issue of spillovers, or other failures of SUTVA. A handful of studies have tried to estimate the incidence spillover effects directly using inter-regional comparisons (e.g., Blundell, Dias, Meghir, and

[10] Similar approaches have been pursued in (e.g., Blundell, Bozio, and Laroque 2012).
[11] It is worth noting that whether spillover effects lead to a bias in the treatment effect depends in part on the outcome and the way it is affected by spillover effects. For example, if the object of interest is the hazard of exiting unemployment, then a standard matching function would imply that spillover effects due to crowding or general equilibrium responses in vacancies would affect the hazard and treatment group proportionately. Taking the log of the hazard rates would then lead to unbiased estimates.

Van Reenen 2004, Ferracci, Jolivet, and van den Berg 2010, Gautier, Muller, van der Klaauw, Rosholm, and Svarer 2011). There are roughly two approaches, neither of which is able to fully identify the spillover effect. One approach is to compare the outcome of control group to similar individuals in untreated areas. Another approach is to compare effect of treatment across sites with different treatment intensity or labor market conditions. The difficulty hereby is that typically neither the treatment site nor the size of the treatment group (and hence the amount of potential spillover) is randomly assigned. For example, Hotz (XXX) discusses that sites for the JTPA evaluations were not selected randomly. Alcott (2013) studies the sources of observed bias in site-selection in a large electricity conservation experiment. In Section V we discuss a recent paper by Crepon, Duflo, Gurgand, Rathelot, and Zamora (2013) that resolves this problem in the context of a job search assistance program by randomly assigning both the treatment and the number of workers treated.

Absent such a multi-stage experimental design, relatively few options are available to researchers to assess the degree of the actual or potential spillover effects present in the context of their evaluation. One approach has been to assess the potential degree of spillover effects via simulations. Thereby, a difficulty is that the degree of the spillover effect crucially depends on the response in job creation to changes in labor supply. An area of research where spillover effects have received substantial attention is the analysis of the employment and welfare impacts of extensions in unemployment insurance benefits. To assess the potential degree of spillovers, several papers have used estimates of the matching function to adjust

micro-econometric estimates of the effect of policy-induced changes in unemployment insurance durations on unemployment duration or exit hazards for the presence of crowding.[12] Such ad-hoc simulations are partial-equilibrium in nature, and could be interpreted as a short-run effect, when vacancies have not yet adjusted. To take into account vacancy responses, Landais, Michaillat, and Saez (2012) specify a general equilibrium model of the labor market that nests two scenarios. In a standard, competitive search-matching model, the vacancy response to changes in labor supply is sufficiently strong to offset the crowding effect completely. If jobs are rationed, on the other hand, increased competition by subsidized job seekers leads to crowding even in general equilibrium. In principle, a calibrated model of this kind could be used to assess the potential degree of crowding. However, the ongoing debate on the general equilibrium effect of UI extensions shows the difficulty of choosing the appropriate parameterization of employer responses.

In the spirit of using random variation in the treatment across localities to assess the presence of spillover effects, a couple of papers have tried to exploit region-specific changes in policy-induced UI variation in the U.S. to assess the full effect of the policy on the entire labor market (Hagedorn, Karahan, Manovskii, and Mitman 2015, Hagedorn, Manovskii, and Mitman 2015). Since UI variations usually depend on economic conditions at the state level, these studies use border

[12] One added difficulty in the case of UI is that in most cases in the U.S. the policy-induced changes in the level or duration of UI benefits are a function of labor market conditions – making it crucial to properly control for the direct effect of local labor market conditions.

communities unaffected by the policy change as counterfactual.[13] A potential difficulty stressed in the literature on evaluation of place-based policies is the presence of spatial spillovers between adjacent or related labor market areas. Such spillovers again constitute a failure of SUTVA, and can lead to biased estimate of region-level policy effects. Cerqua and Pellegrini (2014) develop alternative estimates to the TOT that take into account the degree of spatial spillover effects.

Another source of SUTVA failures are interactions between treatment and control participants. Such 'dilution' effects can lead to an underestimation of the treatment effect. If possible, a typical approach to circumvent such interactions is to raise the level of randomization (say, from a sub-group within a site to a whole site). This approach can help to avoid interactions between individuals in the treatment and control groups. It does not resolve potential interactions between treated participants. This may be part of the mechanism of the treatment; it may also be a potentially unintended variation in treatment intensity that we discuss under site effects. In either case, when designing and experiment, it might be valuable to consider ways of keeping track of social interactions, perhaps by asking about friends in a baseline survey, or monitoring (or manipulating) the use of certain kinds of social media.

[13] A key practical difficulty there is that measures of unemployment rates at the sub-state level is often very noisy. Estimates using administrative employment data based on the universe of private employees appear to show little spillover effects (#CITES).

c. Subgroup and site effects

A common question in program evaluation is whether program impacts vary across subgroups defined by observable characteristics (e.g., race, gender, past work experience). On its face, it is straightforward to answer this question in the context of a randomized trial: One simply constructs treatment-control contrasts separately for each subgroup. Many authors emphasize the importance of conducting the randomization separately for each subgroup of interest. This is not in principle necessary – unconditional random assignment ensures that assignment is random conditional on predetermined characteristics as well – but can add power for subgroup comparisons, especially in smaller samples.

A more important issue is the potential number of comparisons to be estimated. Even a program that has no effect on anyone will be likely to show a statistically significant effect for some subgroup, if enough subgroup estimates are computed. (A similar problem arises when considering effects on multiple outcomes.) Researchers have taken a number of approaches to this multiple comparisons problem. One is to specify the subgroups that will be considered, and the hypotheses of interest, before analyzing the data. This can limit the scope for unconscious data mining, and ensures that the number of comparisons that were considered is known. A second is to adjust the p-values of simple treatment-control contrasts for the multiplicity of the comparisons being estimated. An appropriate adjustment makes it possible to obtain accurate p-values for the test of whether the program had any effect on any subgroup. But two issues remain: These tests typically have *very* low power. In addition, even when they do reject they are often

not able to identify *which* subgroups have non-zero treatment effects. A full discussion of adjustment for multiple comparisons is beyond the scope of this chapter, but Anderson (#CITE) is a useful reference.

One type of subgroup analysis that is quite common is the analysis of treatment effect variation across sites and/or providers. As discussed in Section 3, there are many reasons to expect and be interested in measuring such variation. But with many experimental sites, the number of comparisons can be large, and the available samples for each site can be small. Moreover, the fact that the site-specific treatment effects can in some sense be seen as draws from a larger distribution opens up new options for analysis that are not available in traditional studies of subgroup treatment effects.

The mid-1990s National Job Corps Study illustrates some of the issues involved.[14] Job Corps is a $1.5 billion vocational education and training program that targets disadvantaged youths. The random-assignment study indicated that the program has a positive effect on earnings four years after participation, of a magnitude roughly comparable to the return to a full year of education (Schochet, Burghardt, and McConnell 2008). (At the time of the evaluation, the average participant was enrolled for about eight months.)

But like other job training programs, the specific "treatment" provided to Job Corps participants varies substantially across individuals, according to perceived needs. Moreover, Job Corps services are delivered at 110 mostly residential centers, the majority of which are operated by private contractors. Some providers may

[14] Other studies that examine similar questions are #Bloom study of MTO# and Barnow (2000).

better at delivering an effective program (or at guiding participants to the types of services that they need) than are others. The center-specific treatment effects are thus of great interest.

The Department of Labor (DOL) has long used a performance measurement system to track performance of the different centers and inform decisions about contract renewal. Performance measures are non-experimental, and include statistics like the GED attainment rate or average full-time employment rate of program participants at each center. But it is not clear that these performance indicators successfully distinguish center impacts from differences in the populations served by the various centers.

Schochet and Burghardt (2008; hereafter "SB") attempt to use the random-assignment Job Corps Study to validate DOL's performance indicators (see also Barnow, 2000, who carries out a similar exercise for JTPA). In principle, estimation of site-level causal effects using the experiment is straightforward: One simply compares mean outcomes of the treatment and control groups at each site, relying on the overall random assignment to ensure balance of each site-level comparison. But a few challenges arise.

First, in the Job Corps Study randomization took place before applicants were assigned to centers. Thus, treated individuals are associated with centers, but control individuals are not. SB address this by using intake counselors' assessments of the center that the applicant would most likely attend, collected prior to randomization. To ensure that treatment and control individuals are treated comparably, they use this prediction for both groups, even when it differs from the

actual treatment assignment. Differences occurred for only 7 percent of treatment group enrollees, largely because participants tend to enroll in the closest center or in one that offers a particular vocational program.

Second, even a large RCT sample – the Job Corps Study included over 15,000 participants – can have very small sample sizes at the individual site level. Rather than estimate center-specific treatment effects, SB divide centers into three groups based on their non-experimental performance measures and estimate mean treatment effects for each group. Interestingly, they find that mean program impacts do not differ significantly across groups, suggesting that the performance measurement system is not successfully identifying variation in centers' causal impacts. A related exercise is carried out by Bloom, Hill, and Riccio (2005), who first estimate statistically significant variation in treatment effects across 59 local offices that participated in three welfare-to-work experiments, then use a multi-level model to estimate the relationship between office characteristics – mostly having to do with the way that the treatment was implemented in each site, though they also include the local unemployment rate – and office-level treatment effects. In contrast to the Job Corps study, they do find significant associations of the treatment effect with both their implementation measures and the local unemployment rate.

Bloom, Hill, and Riccio's (2005) interest is in identifying which program features are most effective. It is important to emphasize, however, that the association between site-level characteristics X_j and the site-specific treatment effect τ_j is observational, not experimental, and does not bear a strong causal interpretation. It is quite possible that what appears, for example, to be a strong

association between the emphasis that sites place on quick job placement and the site-level treatment effect instead reflects a non-random distribution of this emphasis across sites that vary in other important ways.

Like the Job Corps study, Bloom et al. (#CITE) do not investigate variation in site impacts conditional on X_j. In many settings, that variation might be of substantial interest. One might like, for example, to estimate effects of individual sites, or to ask which of a number of available performance measures do the best job of predicting experimental impacts. The latter question is a natural one to ask regarding the Job Corps Study, but to our knowledge it has not been studied [#DOL contracted this out in 2010-I wonder if anything came of it?].

Much work on the estimation of site effects themselves comes out of efforts to measure of hospital, school, or teacher performance (see, e.g., #), often with non-experimental data. As in the Job Corps Study, samples are frequently small at the site level, so site-specific treatment effect estimates are quite noisy. One consequence is that actual treatment effects will typically be closer to the average than are estimated effects, even when the research design permits unbiased estimation of each effect. Thus, it is common in these literatures to "shrink" the estimated treatment effects toward the mean. The procedure goes by many different names – e.g., shrinkage, Empirical Bayes, regularization, partial pooling, multi-level modeling – but the basic idea is that the posterior estimate of a site's effect equals a weighted average of the unbiased estimate of that site's effect and the mean site effect, with weights that depend on the precision of the site estimate. Let τ_j represent the impact of the program at site j, and suppose that across sites, $\tau_j \sim$

$N(\tau bar, \omega^2)$. Suppose that we have a noisy but unbiased estimate of the site j effect: $t_j \mid \tau_j \sim N(\tau_j, \sigma^2)$. Then the former can be treated as a prior distribution for τ_j. By Bayes' Rule, the posterior mean of τ_j is

$$E[\tau_j \mid t_j] = (1\text{-}f)\, \tau bar + f\,(t_j - \tau bar),$$

where

$$f = \omega^2 / (\omega^2 + \sigma^2)$$

is the reliability ratio of the site-specific effect estimate.

When the treatment effect varies systematically with site-level covariates – characteristics either of the treatment or of the counterfactual – this can be used to improve precision. If the site effects are modeled as a function of site characteristics, $\tau_j = X_j \beta + \nu_j$, with $\nu_j \sim N(0, \sigma_\nu^2)$, then the noisy site-level estimate t_j should be shrunken toward the conditional mean rather than to the grand mean:

$$E[\tau_j \mid t_j, X_j] = (1\text{-}f')\, X_j\beta + f'\,(t_j - X_j\beta),$$

where f' is the conditional reliability ratio, $f = \omega^2 / (\omega^2 + \sigma_\nu^2)$. This is sometimes known in the statistics literature as "partial pooling."

One use of the shrinkage approach is by Kane and Staiger (2008; see also #), who use a random-assignment experiment to validate non-experimental estimates of teachers' treatment effects on their students. They shrink the non-experimental estimates to construct posterior means, under the assumption that these estimates are valid, and ask whether the result is an unbiased predictor of a teacher's treatment effects under random assignment.

Kane and Staiger focus on "value-added" scores, regression estimates of teachers' effects on their students' test scores from observational regressions, as the

sole non-experimental estimate. They fail to reject the hypothesis that these scores are unbiased predictors of the experimental effects, consistent with the view that they are unconfounded by student sorting. But the experiment has quite low power to distinguish alternative explanations, and Rothstein (2014) argues that the question remains unresolved.

A natural follow-up question is whether other non-experimental measures (e.g., classroom observations) can improve the prediction of experimental effects. If so, one might want to use a weighted average of the available measures, weighted to best predict the experimental treatment effect, for performance measurement purposes. To our knowledge, no study has attempted to estimate these weights.

d. Heterogeneous Treatment Effects and External Validity

The empirical literature on program evaluation has been increasingly aware of the role of heterogeneity in treatment effects for interpreting estimates of program impacts and assessing their external validity. One can roughly distinguish between two sources of heterogeneity. First, treatment effects may vary because of differences in characteristics at the individual level (such as preferences, abilities, health, beliefs, resources, family environment, or access to networks), differences in characteristics of the environment (such as state of the labor market, including business cycle and industry or occupation structure, population density, or labor market discrimination), and differences in aspects of the program (such as unintended differences in the intensity of treatment, something we address under site effects). Second, treatment effects may also vary because of variation in

structural aspects of the program, such as differences in work incentives. In this context, the literature has focused mainly on assessing the effect of heterogeneity for understanding program impacts. Assessing heterogeneity in program impacts is clearly important for understanding the nature of the program, to do cost-benefit analysis, or to assess potential interactions with other program serving certain affected sub-populations. Fewer studies have addressed the question to what extent presence of heterogeneity affects our ability to use estimated program impacts to predict the effect of the same program in other populations or environments.

The literature is broadly in agreement on how to deal with heterogeneity in treatment effects by *observable* characteristics of study participants. As discussed in Section IV.c, the experimental design implies that one can obtain consistent estimates of the treatment impact for each subgroup, subject to having sufficiently large sample sizes. One can then extrapolate the TOT and ATE to settings with other distribution of observable characteristics by constructing appropriately weighted averages of subgroup effects and corresponding standard errors. As a more common alternative, one can directly estimate TOT and ATE for another population by reweighting the original sample to match the distribution of observable characteristics of the alternative population (e.g., DiNardo, Fortin, and Lemieux 1997). If multiple treatment sites are available, in principle a similar approach can be used to assess the effect of environmental characteristics, such as labor market conditions or industrial structure.

The case of heterogeneity by *unobserved* characteristics has presented greater challenges. It is widely accepted that in the case of imperfect compliance,

under assumption of monotonic responses to the inducement to take up the program, the estimated treatment effect captures the local average treatment effect (LATE) for the compliers (Imbens and Angrist 1994, Angrist, Imbens, and Rubin 1996). But with heterogeneous treatment effects, even under perfect compliance the TOT for a particular treatment population may not be relevant for other populations of interest. A key question then is how representative the LATE is for the group of people that would be potentially affected by the program in question. In many cases the program compliers are likely to be similar, in which case LATE is the relevant parameter. In other cases – for example when compliance is likely to differ – the estimated LATE from one program evaluation may be less useful.

Heckman and Vytlacil (2005) propose a conceptual framework to analyze heterogeneity in treatment effects that relies on the concept of marginal treatment effect (MTE). If τ_i denotes the individual treatment effect, X_i is a vector of observed individual characteristics, and v_i is the error in the equation determining take up of treatment, then the marginal treatment effect is defined as $E[\tau_i | X_i = x, v_i=v]$. For practical purposes, the MTE is useful because it can be calculated when a multi-valued instrument for program participation is available. In that case, the MTE is the marginal effect of the outcome on the predicted probability of take up. In other words, the MTE can be obtained by a non-parametric regression of the outcome on the fitted probability of program participation resulting from the first stage equation.[15]

[15] Many other relevant parameters, including LATE and ATE, can be expressed as functions of the MTE. However, to estimate the ATE or the TOT, say, one needs to obtain the MTE for each value of X

This is not possible in the case of a simple RCT. However, when the RCT is taking place at multiple sites, the relationship between the site-specific compliance rate and the site-specific treatment effect (i.e., the site-specific LATE) could be use to estimate a relationship between the compliance rate and the treatment effect.[16] (Alternatively, one could directly regress the site-specific mean outcome on the estimated probability of take up and obtain the MTE for different compliance rates.) This relationship could in principle be used to forecast the local average treatment effect at a potential alternative treatment site (possibly reweighting to adjust for differences in observable characteristics). To do so, one could first predict the compliance rate at the alternative treatment site based on observable characteristics. More generally, this approach would allow inferring the effect of any intervention affecting the cost of compliance and hence the compliance rate itself.

A closely related question is under what circumstances on can draw inferences about the distribution of treatment effects. As mentioned at the outset, knowledge of the distribution of heterogeneous treatment effects is undoubtedly important in assessing the impact of a particular program. However, it is less straightforward how such information can be use to address the issue of external validity if treatment effects vary purely with unobserved characteristics.

One approach that has been used to make inferences about heterogeneity in

---

for the full range of complier probabilities, i.e., from 0 to 1. While in many cases this may be infeasible due to data limitations, if available this could be used to extrapolate the ATE or TOT for populations with different compliance rates and distribution of characteristics.

[16] Note that the weighting function of the LATE estimator for multi-valued instruments in Imbens and Angrist (1994) is proportional to the differences in take up probabilities between different values of the instrument (ordered by the values' impact on take up). This difference can be interpreted as the difference in compliance between instrument values.

treatment effects is estimation of quantile treatment effects (QTE). As discussed in Section IV.a, QTE for the q-th quantile is defined as the difference in the q-th quantile of the outcome distribution in the treatment and control groups, respectively. It is clear that absent strong assumptions, such as rank stability, QTE do not recover the distribution of treatment effects. Yet, it can be a helpful and easy to implement diagnostic device in at least two senses. First, QTE can be used to reject the assumption of constant treatment effects, which would imply that the QTE is equal at all quantiles. Second, in some cases particular features of a program allows one to derive predictions as to responses in different quantiles of the outcome distribution (see below). More generally, QTE may provide a broad descriptive sense of potential treatment responses.

To make inference on the actual distribution of treatment effects additional assumptions are required. Heckman, Smith, and Clements (1997) show that without such assumptions, the observed experimental data is essentially uninformative about the treatment effects distribution. Moreover, they demonstrate that quite strong assumptions on the dependence of counterfactual outcomes in the control and treatment states are needed to obtain plausible estimates of the distribution of the effect of training in the context of the National Job Training Partnership Act (JTPA) study.

A second type of treatment effect heterogeneity can arise from differences in the structure of the program to be evaluated. In this case, theory may provide weak assumptions that allow making inference on the distribution of treatment effects. Welfare programs represent a good example, since they usually combine a range of

different labor supply incentives arising among others from welfare payments, earnings disregards, implicit tax rates, or phase-out regions. Clearly, these incentives interact locally with individual heterogeneity in preferences or ability, something we will return to below. But the additional structure can make for more natural identifying restrictions than in the case of a program that is at least intended to be uniform, such as a training course. A series of papers has addressed this question in the context of evaluation of Connecticut's welfare-to-work program, Jobs First, against the then-prevailing alternative, the AFDC. For example, to assess the degree of heterogeneity in treatment responses Bitler, Hoynes, and Gelbach (1996) implement a QTE as described above, and relate the resulting estimates to prediction from a standard labor supply model. Kline and Tartari (2014) take this approach one step further and use the prediction from the theory to directly infer about the distribution of treatment effects. As already discussed in Section III.a, given the nature of the program, this effectively means estimating bounds on transition probabilities between a limited set of discrete counterfactual states. Their approach allows Kline and Tartari (2014) to directly test for the prevalence of a range different counterfactual treatment responses to the Jobs First program. Therefore, this is an important diagnostic device for assessing the range of counterfactual treatment responses to the program itself. As discussed above, a potential drawback is that the procedure is rather complex and only applies to the particular program studies. One also has to contend with possibly wide bounds.

In principle, Kline and Tartari's approach can also be used for predicting the effect on the distribution of marginal outcomes of moving from the AFDC to a

welfare-to-work program of the same structure at another site (see Section III.a). Yet, it is worth keeping in mind that the estimated bounds have the LATE property, i.e., they may depend on the particular distribution of individual characteristics and the local environment. Extrapolating to different populations or environments in their context would require imposing additional assumptions on the underlying static labor supply model, and thus trade off additional predictions with robustness.

e. Hidden treatments

A long-standing issue in the evaluation of job training programs is that these evaluations commonly have very high rates of non-compliance and crossovers. Many people assigned to receive training do not complete their courses, and it has been operationally and politically difficult to exclude people assigned to the control group from treatment. Indeed, in some cases, ethical concerns led to decisions to actively inform control group individuals about alternative sources of training.

Much of the literature treats this as non-compliance of the type discussed in Section 3, so estimates the training effect by dividing the ITT effect by an estimate of the complier share (see, e.g., Heckman, Hohmann, Smith, and Koo, 2000). But this is unsatisfactory. In many programs, the training offered to the treatment group differs from that which the control group non-compliers are able to obtain. In technical terms, this is a violation of SUTVA; practically, it means that assignment to treatment may affect outcomes even for the always-takers who receive (some type of) training in any case. To our knowledge, this issue has not been addressed in the enormous literature on job training experiments. (Heckman et al., 2000, note the

issue, but their analyses focus on non-random selection into training and heterogeneity of training effects, which are related but distinct issues.)

A very recent literature takes up this topic of "hidden" treatments in the context of the Head Start pre-school program. The Head Start Impact Study randomly assigned Head Start applicants to be offered care or turned away. Many of the control group applicants (and a smaller share of the treatment group) wound up receiving alternative center-based childcare that is thought to be less effective but may be a partial substitute. Walters (2014) estimates heterogeneity in the Head Start effect across centers (sites), finding (among other results) that the LATE is smaller when more of the complier group is drawn from other centers rather than home-based care.

Kline and Walters (2014) explicitly model the hidden alternative center treatment, using variation in the compliance patterns across participants' observable characteristics (e.g., parental education) to identify a multinomial variant of a Heckman (1979) parametric selection correction and thus obtain partially experimental estimates of the separate effects of the two types of child care. Their approach leverages variation across observable characteristics (X) in the share of experimental compliers who are drawn from alternative center care, together with a utility-maximizing choice model that constrains how selection on *un*observables varies with X. With the restrictions imposed by this model, they find large effects of Head Start relative to home-based care. As the Head Start experiment did not directly manipulate the choice between home-based and other center care, they are not able to estimate the relative effect of these with any precision in their

least restrictive model, though point estimates are consistent with an effect of other centers comparable to that of Head Start. When Kline and Walters impose stronger restrictions on the selection process, they obtain similar point estimates but with more precision.

Feller et al. (2014) also examine the hidden treatments issue in the Head Start Impact Study sample. They use a principal poststratification approach that, like Kline and Walters, exploits variation across observables in selection into the two treatments. They couple this to a finite mixture modeling strategy that treats the separation of the two complier subgroup distributions as a deconvolution exercise. They impose parametric assumptions about these distributions to identify the local average treatment effects of the two treatments. Results are similar to Kline and Walters: Head Start has positive effects on those who would otherwise be at home, but little effect on those who would otherwise receive alternative center-based care. Another example of the analysis of hidden treatments is Pinto's (2015) analysis of the Moving to Opportunity experiment. In one view, the MTO study involved two treatment arms: One offered a housing voucher that could be used anywhere, and the other restricted the voucher to a low-poverty neighborhood. Straightforward experimental comparisons identify the ITT and LATE of usage of each type of voucher. In another view, however, the relevant treatment is the type of neighborhood in which the participant lives. Pinto (2015) uses revealed preference restrictions – anyone offered an unrestricted voucher who moves to a low-poverty neighborhood can be assumed to choose the same type of neighborhood in the counterfactual where she receives a restricted voucher – to identify parameters of

interest concerning the distribution of neighborhood-type treatment effects.

f. Mechanisms and multiple treatments

Researchers have used a number of strategies to extract from experimental data evidence on the mechanisms underlying the treatment effects that the experiments identified.

In the simplest case, it is sometimes possible to use experimental variation to distinguish the relevant mechanisms, with only minimal restrictions derived from theory. This is most feasible when the experiment involves more than two groups. The first large-scale social experiments, the Negative Income Tax studies, were used in this way. The "treatment" here was a tax schedule described by two parameters: The transfer received if earnings were zero and the tax rate applied to any earnings. The main outcome was labor supply, and a key concern of these studies was to distinguish income from substitution effects.

With a single treatment arm and a single control group, this would not be possible: The net effect of the treatment would be identified, but there would be no way of distinguishing substitution from income effects. (One exception would be if the treatment were designed to be a fully compensated change in the marginal tax rate – this would have no income effect, so the treatment effect would equal the substitution effect. But the NIT treatments were not designed this way.) With multiple treatments that vary both the base transfer and the marginal tax rate, and with an assumption that both income and substitution effects are linear in the relevant tax variable, the two can be distinguished.

To see this, suppose a labor supply function that relates hours of work (H) to the wage rate (w), non-labor income (N), the marginal tax rate (r), and other factors such as preferences for leisure (e):

$$H=f(w, N, r, e).$$

For simplicity of exposition, we assume a constant marginal tax rate, though this is not crucial (#Hausman). A more restrictive assumption is that the individual labor supply function is linear and additively separable in non-labor income and the net-of-tax hourly wage:

$$H_i = \gamma_i + w_i(1-r_i)\,\delta_i + N_i\,\eta.$$

Now consider a simple experiment that assigns some individuals to a control group where r_i and N_i are not manipulated and others to a treatment group that receives an additional baseline transfer D and faces an increment to the tax rate t. Then, adopting the earlier potential outcomes framework, each individual has two potential outcomes:

$$H_{i0} = \gamma_i + w_i(1-r_i)\,\delta_i + N_i\,\eta_i \text{ and}$$

$$H_{i1} = \gamma_i + w_i(1-r_i - t)\,\delta_i + (N_i + D)\,\eta_i.$$

With random assignment, the difference in mean labor supply between treatment and control groups equals

$$E[H_i \mid D_i =1] - E[H_i \mid D_i = 0] = -t\,E[w_i\,\delta_i] + D\,E[\,\eta_i\,].$$

The first term here represents substitution effects, while the second represents income effects. But the simple experiment identifies only the combination of them.

Fortunately, the NIT studies involved multiple treatment arms, with various combinations of transfers and tax rates. Consider a simple extension of the above structure, with two treatment groups 1 and 2 and associated parameters $\{D_1, t_1\}$ and $\{D_2, t_2\}$. Now each individual has three potential outcomes associated with assignment to the control group and each of the treatment groups, H_0, H_1, and H_2. Two distinct treatment-control contrasts can be computed:

$$E[H_i \mid D_i = 1] - E[H_i \mid D_i = 0] = -t_1 E[w_i \delta_i] + D_1 E[\eta_i] \text{ and}$$

$$E[H_i \mid D_i = 2] - E[H_i \mid D_i = 0] = -t_2 E[w_i \delta_i] + D_2 E[\eta_i].$$

This is a system of two linear equations and two unknowns. So long as the system has full rank – here, as long as $(D_1/D_2 \neq t_1 / t_2)$ – it can be solved for the mean income elasticity of labor supply, $E[\eta_i]$, and for $E[w_i \delta_i]$. The latter can be divided by the mean wage rate, $E[w_i]$, to obtain a wage-rate-weighted mean substitution elasticity. (With a large enough sample, the mean substitution elasticity, $E[\delta_i]$, could be identified by stratifying the treatment-control comparison by the wage rate.)

A number of studies used the NIT experiment data to estimate the parameters of the labor supply function in basically this way, accounting for additional complications that we neglect here (e.g., participation decisions, non-linear tax schedules, etc.) and often using more complex labor supply functions. See, e.g., Moffitt (1979). But this was by no means universal: In the late 1970s, the experimental paradigm was not as well developed, and many of the studies that used the experimental data did not rely solely on the randomly assigned components of non-labor income and tax rates for identification (e.g., Keeley et al., 1978).

In the above simple model the mean income and labor supply elasticities are just identified with two treatment arms. With more than two arms – the Seattle/Denver experiment alone had 11 – the model is over-identified. This opens the possibility of performing over-identification tests of the restrictions imposed when specifying the labor supply function. Ashenfelter and Plant (1990) estimate separate treatment effects of each treatment arm, but we are not aware of studies that investigate formally whether the pattern of effects is consistent with the posited labor supply function.

Card and Hyslop (2005) [henceforth CH] analyze the data from the Canadian Self Sufficiency Program (SSP) RCT. SSP, a welfare-to-work program, combined a strong, temporary work incentive for participating workers with a fixed initial time period during which welfare recipients had to establish eligibility in the program by working full time. As a result of this two-tiered structure, it is difficult to assess the various components of the program from the RCT alone. This makes it difficult to compare the effects of SSP with other welfare-to-work programs, to assess how SSP worked, and what can be learned for similar programs. To separately estimate the effects of the different SSP components, CH use a parametric statistical model to separately identify effect of the different incentives inherent in the SSP program. In contrast to static evaluations of welfare-to-work programs, CH thereby focus on the dynamic labor supply incentives inherent in the program.

Assessing the dynamic effects of the work subsidy for eligible workers separately from the entitlement effect based on the RCT alone is difficult for several reasons. One cannot directly analyze the effect of the subsidy (which in the

following we will refer to as the SSP program) for those who became eligible because of selection in the eligibility decision. Given imperfect compliance, one can estimate the LATE of SSP on total employment or on the fraction employed at any given point in time. Given the nature of the eligibility process, DH can use their model to make inference in the nature of the initial selection. However, potential differential changes in the nature of selection in the treatment and control groups make it impossible to estimate the dynamic responses of hazard rates or wages just based on the RTC.[17] In addition, as in other welfare evaluations, endogenous employment decisions make an analysis of wage outcomes problematic. Another issue is that in the short run the strong work incentive arising from the option value in the eligibility period is potentially confounded with the effect of the subsidy.

To address these difficulties, CH proceed by developing a logistic model with random effects and heterogeneity to estimate a benchmark for welfare transitions in the absence of SSP (i.e., for the control group). This model is then combined with parametric specifications of the treatment effects over different ranges of the program spell, as implied by incentives inherent in SSP. This step includes modeling the participation decision, as well as welfare transitions as function of the SSP subsidy and current and lagged welfare status. A key assumption thereby is that the chosen controls for heterogeneity and the functional form restrictions are sufficient to control for the selection bias introduce by the eligibility window and the dynamic

[17] CH use a standard search theory to model the incentives of SSP, and capture the effect of eligibility and the SSP subsidy on labor supply incentives via their effects on the reservation wage. The search model clarifies that in the presence of heterogeneity, the pool of workers employed at any given point in time may be selected, whether or not there also is sample selection arising from employment decisions (e.g., Ham and Lalonde 1996).

nature of the problem. CH experiment with different specifications of heterogeneity, and provide ample discussion of the goodness of fit of the model. As a result of this exercise, they are able to obtain separate effects of eligibility and SSP. This allows them to simulate the effects of different components of the program and counterfactual policy changes relating to the time path of the subsidy.

The approach and finding in CH suggest that one may not need a structural model to separately identify multiple treatment effects, the dynamic effects of a program, or to simulate the effect of alternative policies. However, an assumption on function form is required, which can be assessed, as well as harder-to-assess assumptions on the form of underlying heterogeneity.

To estimate mechanisms underlying the effect of experimental or policy variation, other papers have used insights from theory to aid identification without estimating a structural model. For example, Schmieder, von Wachter, and Bender (2014) use insights from the standard search model to estimate the effect of unemployment duration on wages. A recurring question in the analysis and evaluation of welfare and unemployment programs has been the effect of employment and unemployment on productivity and wages. If wages rise with employment duration, welfare-to-work programs can lead to sustained labor force participation. In contrast, if longer nonemployment duration reduces wages, and hence the disincentive to work, more generous benefits can lead to a welfare trap. Card and Hyslop (2005) find that increased employment in the course of the Canadian Self-Sufficiency Program did little to increase wages. In contrast, Grogger (2009) finds positive wage impacts of employment in the context of a randomized

evaluation of Florida's welfare-to-work program. Few papers have directly analyzed effect of unemployment duration on wages.[18] The question is difficult for at least two reasons. First, as in case of Card and Hyslop (2005), even in the presence of exogenous variation in incentives at the group level, the type of worker employed at any given point in the unemployment may differ in the treatment and control groups.[19] In other words, it is difficult to find a valid instrument for the duration of unemployment or unemployment spell. A second complication arises because even if such variation was available, a change in wages might both arise because of a change in wage offers, as well as due to a change in reservation wages.

To address these difficulties, Schmieder, von Wachter, and Bender (2014) use the fact that the canonical search model has the strong prediction that forward-looking individuals valuing future unemployment insurance benefits will respond to a benefit extension by raising their reservation wage well before benefit exhaustion. Unless reservation wages do not bind, this implies that in response to extensions in UI durations one should see a rise in observed reemployment wages throughout the spell. Hence, if one finds reemployment wages at different points of the unemployment spells are identical, this implies reservation wages likely had little effect on observed wages. In that case, the effect of an increase in UI durations on wages can only arise from an effect of the rise in nonemployment durations on offered wages. Hence, an exogenous increase in UI durations can be used as an

[18] An exception is Addison and Blackburn (2000), who discuss some of the arising issues. A larger number of papers has addressed the question of duration dependence in unemployment spell, see Kroft, Lange, and Notowidigdo (2013) and references therein.
[19] This bias arises even in the absence of differences in participation.

instrument to estimate the effect of nonemployment duration on wages.[20]

Schmieder et al. (2014) Study this problem in the context of discontinuous variations in unemployment insurance duration in Germany by exact age. While the maximum duration of UI benefits is exogenous, because of selection, the research design does not allow them to estimate the effect of the program on wages or hazard rates at different duration of unemployment. Yet, they show that even in the presence of selection one can estimate an upper bound for the mean upward shift in the wage distribution. The reduced form effects of the UI extensions Schmieder et al. (2014) are positive for average nonemployment durations and negative for average reemployment wages. They also find that reemployment wages are decreasing over the nonemployment spell,[21] and that the path of reemployment wages is essentially unaffected by UI extensions. Hence, they conclude that the main mechanism leading to longer nonemployment durations is a decline in search intensity, not a rise in reservation wages.[22] As a consequence, the negative wage effect of nonemployment durations can be attributed to increases in nonemployment durations. The resulting instrumental variable estimates suggest a significantly negative effect of unemployment duration on wages.[23]

---

[20] The authors argue that their test excludes any affect of the worker's outside option on wages, and hence the findings are not specific to the particular model.
[21] It is well understood that the OLS relationship between wages and unemployment duration conflates declines in wage offers, declines in wage offers and potential selection throughout the spell.
[22] Several papers show that reservation wages do not appear to respond to the duration of the unemployment spell or to UI duration. Feldstein and Poterba (1984) find that reservation wages are roughly equal to the pre-displacement wage and constant over time. This is confirmed by Krueger and Mueller (2014) who also show that reservation wages do not decline at UI exhaustion.
[23] Schmieder et al. (2014) show that the estimate is the local average treatment effect for those workers whose search intensity is affected by the UI duration; given nonemployment durations are affected throughout the spell, the average is taken over both nonemployment durations and individual heterogeneity in skill depreciation rates.

The incentives in the UI system are inherently dynamic, and the typical approach is to model the dynamic responses is in terms of changes in search intensity and reservation wages. Hence, understanding the determinants of these changes is important for designing optimal UI policies. However, the presence of selection over the unemployment spell makes inference difficult even in presence of experimental variation. Incorporation theoretical insights from search theory directly into the estimation process can be helpful in this setting. To study the determinants of search choices, Della Vigna, Lindner, Reizer, and Schmieder (2014) depart from the observation that observed exit hazards from unemployment typically decline at the beginning of a spell, not rise as the standard model would predict. While selection could explain this pattern, they show that a model based on reference-dependence in preference can explain it without resort to an ad hoc specification of heterogeneity. Moreover, their extended model has added testable implications. Among others, in contrast to the standard model, reference-dependence can predict a spike in exit hazards at UI exhaustion.

Della Vigna et al. (2014) examine these implications in the context of a UI reform in Hungary, which changed the time path of UI benefits, while leaving the benefits in the final tier unchanged. Based on a standard program evaluation, they find their results match the qualitative implications of the reference-dependence model quite closely. In particular, they find evidence of both a rise in the hazard in anticipation of benefit reduction, a spike, and then persistence and signs of habituation after a benefit change. The difficulty with assessing the qualitative prediction alone is that it is unclear whether a reasonable parameterization of

heterogeneity in the standard model could explain the model as well; moreover, it is important to directly assess the goodness of fit of the new model, and estimate the magnitude of the reference dependence. Hence, they proceed to structurally estimate parameterized versions of the standard and augmented search models using minimum distance. As Card and Hyslop (2005), they fit the model on both the monthly exit hazards before and after the benefit reform, varying the UI parameters in the model accordingly.

Using this approach, not surprisingly, they find that introducing reference dependence allows the model to better fit the pattern observed in the data than the standard model, even when heterogeneity is allowed for. The estimation confirms a substantial degree of reference dependence. In principle, the resulting estimates could be used for policy simulations. A potential drawback is that the nature of the data does not allow them to incorporate other forms of heterogeneity that have been shown to affect search behavior, and that may interact with or appear like reference dependence, such as gradual adjustments in spousal income and labor supply, presence of wealth, or savings decisions.

While the potential importance of non-standard components of the utility functions, such as stigma or the utility cost of lying about earnings, have long been incorporated of theoretical or structural models of labor supply and welfare participation, only recently have papers begun to directly estimate such features. Babcock, Congdon, Katz, and Mullainathan (2012) give an overview of the potential importance of behavioral assumption for the evaluation of public programs. Besides Della Vigna et al. (2014), a growing number of papers assesses non-standard or

"behavioral" hypotheses in the context of randomized trials or program evaluation. For example, Lemieux and MacLeod (2000) assess a related form of habituation in the aftermath of dramatic extensions of the Canadian unemployment insurance system. They claim that when a UI system becomes more generous, the effect on the population accrues gradually, as an increasing number of individuals are exposed to and learn about the new system and social norms change. Their findings are broadly consistent with the view that stigma for being on welfare or UI may decline with increasing exposure to the system. Dahl, Kostol, and Mostad (2014) find that random parental exposure to disability insurance in Norway raises the children's propensity to be on disability insurance as well. While this effect is again consistent with stigma being endogenous, it could partly be explained by learning about the availability and accessibility of the program. Other behavioral aspects are likely to affect labor supply or human capital investment decisions as well. For example, non-standard discounting has been shown to affect job search (Della Vigna and Paserman 2005) and high school drop out behavior (Oreopoulos 2007). More recently, Chan (2014) examines the role of time-inconsistency in the context of the randomized evaluation of Florida Transition Program.

Most approaches mentioned so far involve inherent uncertainty arising by the modeling assumptions. An approach that incorporates some of that uncertainty is to assessing the role of mechanisms or the effect of multiple treatments is to provide bounds for the various effects. For example, Schmieder et al. (2014) use their model and their empirical estimates of the shift in the reemployment wage path to provide an upper bound of the potentially confounding effect of reservation

wage changes. Card and Hyslop (2005) and Della Vigna et al. (2014) provide sensitivity analysis, but stop short of deriving bounds. Kline and Tartari (2014), already mentioned in Section 5.a, assess what can be learned without any functional for assumptions in the context of the randomized evaluation of Connecticut's welfare-to-work program, Jobs First. A difficulty with the studies we discussed so far is that the focus on identification and the use of experimental variation has kept the models relatively simple. Hence, their potential use for counterfactual policy simulations is limited. Another approach is to try to use the experimental variation to help identify a richer structural model, as further discussed below.

A closely related topic to the question of mechanisms is the extrapolation of experimental evidence to consider the impacts of new policies, not included in the original evaluation. The value of such extrapolations has long been one of the primary arguments in favor of structural modeling (and against reliance on purely experimental evidence), but some scholars have found out ways to synthesize the approaches. The main challenge here is to bridge between the relatively few parameters that are cleanly identified by an experiment and the larger set of parameters that are needed to characterize most structural models.

One way to do this is to start with a simple characterization of structural behavior that is simple enough to be captured within the experimental evidence. For example, if one assumes that the labor supply function is characterized by constant income and (compensated) substitution elasticities, then the estimates of these parameters that are identified by the NIT experiments are sufficient to identify the effects of alternative NIT parameters that were not included in the experimental

treatments. This sort of exercise is on more solid ground when trying to interpolate to values within the range of tax parameters included in the experiment than when these parameters need to be extrapolated outside of that range, of course.

A more recent, closely related approach is known as the "sufficient statistics" approach (Chetty 2009). Here, the goal is to characterize optimal policy. Starting with a fully characterized (but usually not overly complex) structural model, it is often possible to derive expressions for social welfare, or for the optimal policy, that depend only on a small number of reduced-form parameters. For example, the Baily-Chetty (Baily 1978, Chetty 2006) formula for optimal unemployment insurance benefits expresses the optimal benefit level in terms of the elasticity of unemployment duration with respect to UI benefits, and the income and substitution effects on the exit hazard from unemployment. If one had experimental evidence regarding these effects, one could use the formula to derive the optimal policy (e.g., Chetty 2008, Card, Chetty, and Weber 2007).

Of course, any sufficient statistics approach is dependent upon the validity of the underlying structural model – there is no assurance that the true structural model generates the same sufficient statistics as does the one posited by the researcher. Sufficient statistics may be robust to some violations of the model, so long as they don't change the sufficient statistics. For example, Chetty (2009) gives the example of heterogeneity in treatment effects, where the optimal policy depends only on the mean effect. Yet, it can be hard to know which assumptions in the structural model matter. At a practical level, the optimal policy conclusions may involve extrapolating very far from the range of policy variation included in the

experiment, which means that despite reliance on credibly identified sufficient statistics, one relies crucially on the validity of the theoretical model. In this context, a potential drawback of sufficient statistics is that in contrast to explicitly structural work the empirical fit of the model against the data cannot be assessed.

An alternative approach to obtain a framework for policy extrapolation based on experimental variation is to estimate, or calibrate, a full structural model, using experimental evidence to aid in identifying (some of) the necessary parameters. There are several ways this can be accomplished. First, one can fix individual parameters at the values indicated by experiments, then calibrate or structurally estimate the remainder. This approach is pursued, for example, by Davidson and Woodbury (JPubEc 1997), who use the Illinois reemployment bonus experiment to estimate the parameters of a search cost function, then combine this function with calibrated values, derived from non-experimental data, for other parameters of their model of optimal UI benefits. Second, one can use experimental data to fit a full structural model, but keep the model sufficiently simple such that the main parameters of the model are identified by the available variation, as in DellaVigna, Lindner, Reizer, and Schmieder (2014) discussed above. A draw back of such an approach is that the range of policies that can be examined is limited. The approach can be extended, of course, to estimate a more complex structural model that either relies on additional statistical and theoretical assumptions, additional non-experimental moments, or both. A related strategy is to rely entirely on non-experimental data to estimate a structural model, but to then use experimental

evidence to validate predictions that the model makes for particular reduced-form comparisions (e.g., Todd and Wolpin 2006).

Finally, another approach is to use experimental variation in the incentive to take up a program to effectively estimate a structural model of the compliance rate (e.g., Heckman and Vytlacil 2005). Many economic models of program choice can be expressed as a variation of a classic Roy model (e.g., French and Taber 2014). As described in Section 5.d, Heckman and Vytlacil (2005) show that in case of a multi-valued treatment (instrument), the estimated compliance rate – quasi the fitted values from a first stage regression – can be used to estimate the marginal treatment effect of those individuals induced to participate at each marginal change in compliance. Under suitable assumptions, the resulting relationship between MTE and estimated compliance rate can be used to extrapolate from a given evaluation to any situation where the potential compliance rate can be estimated.

V. Experimental design choices to address potential design issues

In the previous section, we discussed studies that use experimental data, ex post, to address questions of interest that extend beyond the simple mean effect of the treatment on the treated. In most of the cases discussed, these studies were conducted ex post, after the experiment itself was complete. But in some cases it is possible to design the experiment to facilitate examination of the broader questions of interest. This opens up new possibilities and can permit more credible answers than are possible in purely ex post analyses. This section discusses several types of

modifications to the traditional experimental paradigm that can be of particular value.

a. Improved data collection

- *Pre-randomization:*
  - *People have gotten the message that we need good baseline measures.*
  - *But there may be scope for ex ante theorizing (and empirical analysis) about likely dimensions of treatment effect heterogeneity. Important to collect measures of these dimensions.*
  - *Pre-registration of planned heterogeneity analysis and hypotheses. Or big data methods for avoiding false discovery?*
- *Post randomization:*
  - *Focus on measuring the treatment actually received – variation in treatment intensity, hidden treatments, site variation.*
- *Administrative data may help with either.*
  - *Saves money on data collection (e.g., Ashenfelter-Plant argument that use of survey data in NIT study both cost lots of money and compromised validity).*

b. Stratified randomization in relevant dimensions of heterogeneity

- *E.g., stratify by availability of likely alternative treatments, by compliance propensity, or by anticipated treatment effects.*
- *Likely to be limited by practical considerations and low power.*
- *Again, important to plan ahead / preregister / guard against false discovery*

c. Multiple and cross-classified treatments

- *Multiple treatments (e.g., NIT) can help identify model structure.*

- *Cross-classification can help with endogenously observed outcomes, if one of the "treatments" affects observability but doesn't otherwise affect the outcome.*
- *Cross-classification can also be helpful for mechanisms.*
- *Cross-classified designs don't necessarily sacrifice any power.*
    - *Can look at each dimension separately. Estimated effect is an average of LATEs for different values of the other dimensions. But if effects aren't interactive, it doesn't matter. Moreover, averaging the different LATEs isn't bad if the other dimensions reflect variation that arises in the setting of interest.*
    - *One way to achieve cross-classification: Build experiments on top of existing natural experiments.*
- *A particularly useful potential cross-classified treatment: Extra encouragement to comply with experimental assignment, to see how LATEs vary.*
- *Site variation as an example – Crepon, Duflo et al. randomly assign both group-level and individual treatment, to get at contextual/crowdout/general equilibrium effects.*
- *MTO random variation in follow-up survey effort also fits in here.*

d. Incorporation of a relevant theoretical model into the experimental design

- *Hinges, of course, on the model being correct.*
- *Examples:*
    - *Given a structural model, predict TEs, and use these to stratify or target an experimental intervention. (Can see the NIT studies as a simple version of this – treatment probabilities depended on prior earnings.)*
    - *Identify substitution effect via compensated wage changes.*
    - *Vary incentive to participate, to trace out the marginal treatment effect curve (discussed above, but theoretical considerations can tell us how best to vary compliance rate – if we vary the cost in a naturalistic way, we can trace out the MTE curve).*
    - *Treatments designed to get at intertemporal choice. E.g., a reemployment bonus that declines over time, or one that is*

available only to those who survive to a specified point. SSP can be seen as a real example.

VI. Practical considerations in conducting labor market experiments

a. *Access to sampling frame*
b. *Access to program (including right to exclude)*
c. *Cost of implementation*
d. *Feasibility of randomization (including multiple treatments)*
e. *Compliance/SUTVA*
f. *Data collection*
g. *Attrition*
h. *The need for selection models*
i. *Making sure that local effects are interesting*

VII. A fuller overview of labor market experiments

a. *Organize discussion by subject*
b. *Highlight examples that combine theory with randomization.*
c. *Trends:*
    a. *Movement overseas*
    b. *More personnel-style experiments, rather than large scale programs (though ALMPs are an exception to this)*
    c. *More academic-led, rather than by government contracts to large evaluation firms.*
    d. *Internet based experiments*

VIII. Conclusion

References

Addison, J. T., Blackburn, M. L. 2000. The effects of unemployment insurance on postunemployment earnings. Labour Economics, 7(1), 21-53.

Ahn, H., Powell, J. L. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics, 58(1), 3-29.

Allcott, H. 2013. Is Replication Enough? Site Selection Bias in Program Evaluation.

Altonji, J. G., Blank, R. M. 1999. Race and gender in the labor market. Handbook of labor economics, 3, 3143-3259.

Angrist, J. D., Imbens, G. W. 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. Journal of the American Statistical Association, 431-442.

Angrist, J. D., Imbens, G. W., Rubin, D. B. 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434), 444-455.

Babcock, L., Congdon, W. J., Katz, L. F., Mullainathan, S. 2012. Notes on Behavioral Economics and Labor Market Policy. IZA Journal of Labor Policy,1(1), 1-14.

Baily, M. N. 1978. Some aspects of optimal unemployment insurance. Journal of Public Economics, 10(3), 379-402.

Barnow, B. S. 2000. Exploring the relationship between performance management and program impact: A case study of the Job Training Partnership Act. Journal of Policy Analysis and Management, 19(1), 118-141.

Bertrand, M., Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review, 94(4), 991-1013.

Bitler, M. P., Gelbach, J. B., Hoynes, H. W. 2006. What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. The American Economic Review, 988-1012.

Bloom, H. S. (Ed.). 2005. Learning more from social experiments: Evolving analytic approaches. Russell Sage Foundation.

Blundell, R., Bozio, A., Laroque, G. 2011. Labor Supply and the Extensive Margin. The American Economic Review, 101(3), 482-486.

Blundell, R., Meghir, C., Dias, M. C., Reenen, J. V. 2004. Evaluating the Employment Impact of a Mandatory Job Search Program. Journal of the European Economic Association, 2(4), 569-606.

Buchinsky, M. 1994. Changes in the US wage structure 1963-1987: Application of quantile regression. Econometrica: Journal of the Econometric Society, 405-458.

Card, D., Hyslop, D. R. 2005. Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers. Econometrica, 73(6), 1723-1770.

Card, D., Chetty, R., Weber, A. 2007. Cash-on-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market. The Quarterly Journal of Economics, 122(4), 1511-1560.

Cerqua, A., Pellegrini, G. 2014. Do subsidies to private capital boost firms' growth? A multiple regression discontinuity design approach. Journal of Public Economics, 109, 114-126.

Chan, M. K. 2014. Welfare Dependence and Self-Control: An Empirical Analysis (No. 19). Economics Discipline Group, UTS Business School, University of Technology, Sydney.

Chetty, R. 2006. A general formula for the optimal level of social insurance.Journal of Public Economics, 90(10), 1879-1901.

Chetty, R. 2008. Moral Hazard versus Liquidity and Optimal Unemployment Insurance. Journal of Political Economy, 116(2), 173-234.

Chetty, R. 2009. Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance. American Economic Journal: Economic Policy, 1(2), 31-52.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., Zamora, P. 2013. Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. The Quarterly Journal of Economics, 128(2), 531-580.

Dahl, G. B., Kostol, A. R., Mogstad, M. 2013. Family welfare cultures (No. w19237). National Bureau of Economic Research.

Davidson, C., Woodbury, S. A. 1997. Optimal unemployment insurance. Journal of Public Economics, 64(3), 359-387.

DellaVigna, S., Paserman, M. D. 2005. Job Search and Impatience. Journal of Labor Economics, 23(3).

DellaVigna, S., Lindner, A., Reizer, B., Schmieder, J. F. 2014. Reference-Dependent Job Search: Evidence from Hungary. Unpublished Working Paper.

DiNardo, J., Fortin, N. M., Lemieux, T. 1995. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach (No. w5093). National Bureau of Economic Research.

Feldstein, M., Poterba, J. 1984. Unemployment insurance and reservation wages. Journal of Public Economics, 23(1), 141-167.

Feller, Avi and Grindal, Todd and Miratrix, Luke W. and Page, Lindsay C. 2014. Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings. Available at SSRN: http://ssrn.com/abstract=2534811

Ferracci, M., Jolivet, G., van den Berg, G. J. 2010. Treatment Evaluation in the Case of Interactions within Markets (No. 4700). Institute for the Study of Labor (IZA).

Fraker, T., Maynard, R. 1987. The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, 194-227.

French, E., Taber, C. 2011. Identification of Models of the Labor Market. Handbook of Labor Economics, 4, 537-617.

Gautier, P. A., Muller, P., Rosholm, M., Svarer, M., van der Klaauw, B. 2012. Estimating Equilibrium Effects of Job Search Assistance (No. 9066). CEPR Discussion Papers.

Grogger, J. 2009. Welfare reform, returns to experience, and wages: using reservation wages to account for sample selection bias. The Review of Economics and Statistics, 91(3), 490-502.

Gronau, R. 1974. The effect of children on the housewife's value of time. In Economics of the family: Marriage, children, and human capital (pp. 457-490). UMI.

Hagedorn, M., Karahan, F., Manovskii, I., Mitman, K. 2013. Unemployment Benefits and Unemployment in the Great Recession: the Role of Macro Effects (No. w19499). National Bureau of Economic Research.

Hagedorn, M., Manovskii, I., Mitman, K. 2015. The Impact of Unemployment Benefit Extensions on Employment: The 2014 Employment Miracle? (No. w20884). National Bureau of Economic Research.

Ham, J. C., Li, X., Reagan, P. B. 2011. Matching and semi-parametric IV estimation, a distance-based measure of migration, and the wages of young men. Journal of Econometrics, 161(2), 208-227.

Hausman, J. A., Wise, D. A. 1979. Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. Econometrica, 47(2), 455-73.

Heckman, J. J. 1979. Sample Selection Bias as a Specification Error. Econometrica, 47(1), 153-61.

Heckman, J. J., Hotz, V. J. 1989. Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. Journal of the American statistical Association, 84(408), 862-874.

Heckman, J. J., Smith, J. 1997. Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. Review of Economic Studies, 64(4), 487-535.

Heckman, J. J., Vytlacil, E. 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica, 73(3), 669-738.

Heckman, J., Hohmann, N., Smith, J., Khoo, M. 2000. Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. The Quarterly Journal of Economics, 115(2), 651-694.

Holland, P. W. 1986. Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), 945-960.

Horowitz, J. L., Manski, C. F. 2000. Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. Journal of the American Statistical Association, 95(449), 77-84.

Johnson, W., Kitamura, Y., Neal, D. 2000. Evaluating a Simple Method for Estimating Black-White Gaps in Median Wages. American Economic Review, 90(2), 339-343.

Kane, T. J., Staiger, D. O. 2008. Estimating teacher impacts on student achievement: An experimental evaluation (No. w14607). National Bureau of Economic Research.

Kline, P., Tartari, M. 2015. Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach. NBER Working Paper, (w20838).

Kline, P., Walters, C. 2014. Evaluating Public Programs with Close Substitutes: The Case of Head start. IRLE Working Paper #123-14.

Kling, J. R., Liebman, J. B., Katz, L. F. 2007. Experimental analysis of neighborhood effects. Econometrica, 75(1), 83-119.

Kroft, K., Lange, F., Notowidigdo, M. J. 2013. Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment. The Quarterly Journal of Economics, 128(3), 1123-1167.

Krueger, A. B., Mueller, A. I. 2014. A Contribution to the Empirics of Reservation Wages (No. w19870). National Bureau of Economic Research.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604-620.

Landais, C., Michaillat, P., Saez, E. 2010. Optimal Unemployment Insurance over the Business Cycle (No. w16526). National Bureau of Economic Research.

Lee, D. S. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies, 76(3), 1071-1102.

Lemieux, T., MacLeod, W. B. 2000. Supply Side Hysteresis: the Vase of the Canadian Unemployment Insurance System. Journal of Public Economics,78(1), 139-170.

Looney, A., Manoli, D. 2013. Are There Returns to Experience at Low-Skill Jobs? Evidence from Single Mothers in the United States over the 1990s.

Moffitt, R. 1985. Unemployment Insurance and the Distribution of Unemployment Spells. Journal of Econometrics, 28(1), 85-101.

Neal, D. A., Johnson, W. R. 1996. The Role of Premarket Factors in Black-White Wage Differences. The Journal of Political Economy, 104(5), 869-895.

Newey, W., Powell, J. L., Walker, J. R. 1990. Semiparametric Estimation of Selection Models: Some Empirical Results. American Economic Review, 80(2), 324-28.

Oreopoulos, P. 2007. Do Dropouts Drop Out Too Soon? Wealth, Health and Happiness from Compulsory Schooling. Journal of Public Economics, 91, 2213-2229.

Pinto, R. 2015. Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity. Mimeo., University of Chicago.

Powell, J. L. 1984. Least absolute deviations estimation for the censored regression model. Journal of Econometrics, 25(3), 303-325.

Rothstein, J. 2014. Revisiting the impacts of teachers. Unpublished working paper. http://eml. berkeley. edu/~ jrothst/workingpapers/rothstein_cfr. pdf.

Schmieder, J. F., von Wachter, T., Bender, S. 2014. The Causal Effect of Unemployment Duration on Wages: Evidence from Unemployment Insurance Extensions. IZA Discussion Paper No. 8700.

Schochet, P. Z., Burghardt, J. A. 2008. Do Job Corps Performance Measures Track Program Impacts? Journal of Policy Analysis and Management, 27(3), 556-576.

Schochet, P., Burghardt, J., McConnell, S. 2008. Does Job Corps Work? Impact Findings from the National Job Corps Study. The American Economic Review, 98(5), 1864-1886.

Todd, P. E., Wolpin, K. I. 2006. Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility. American Economic Review,96(5), 1384-1417.

Vytlacil, E. 2002. Independence, Monotonicity, and Latent Index Models: An Equivalence Result. Econometrica, 70(1), 331-341.

Walters, C. 2014. Inputs in the production of early childhood human capital: Evidence from Head Start (No. w20639). National Bureau of Economic Research.