NBER WORKING PAPER SERIES

PROGRAM EVALUATION AND RESEARCH DESIGNS

John DiNardo
David S. Lee

Working Paper 16016
http://www.nber.org/papers/w16016

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2010

We are grateful to Diane Alexander and Pauline Leung, who provided outstanding research assistance. We thank Orley Ashenfelter, David Card, Damon Clark, Nicole Fortin, Thomas Lemieux, Enrico Moretti, Phil Oreopolous, Zhuan Pei, Chris Taber, Petra Todd, John Van Reenen, and Ken Wolpin for helpful suggestions, comments, and discussions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Program Evaluation and Research Designs
John DiNardo and David S. Lee
NBER Working Paper No. 16016
May 2010
JEL No. C10,C50,C52,H00,I00,J00,J24

ABSTRACT

This chapter provides a selective review of some contemporary approaches to program evaluation. One motivation for our review is the recent emergence and increasing use of a particular kind of "program" in applied microeconomic research, the so-called Regression Discontinuity (RD) Design of Thistlethwaite and Campbell (1960). We organize our discussion of these various research designs by how they secure internal validity: in this view, the RD design can been seen as a close "cousin" of the randomized experiment. An important distinction which emerges from our discussion of "heterogeneous treatment effects" is between ex post (descriptive) and ex ante (predictive) evaluations; these two types of evaluations have distinct, but complementary goals. A second important distinction we make is between statistical statements that are descriptions of our knowledge of the program assignment process and statistical statements that are structural assumptions about individual behavior. Using these distinctions, we examine some commonly employed evaluation strategies, and assess them with a common set of criteria for "internal validity", the foremost goal of an ex post evaluation. In some cases, we also provide some concrete illustrations of how internally valid causal estimates can be supplemented with specific structural assumptions to address "external validity": the estimate from an internally valid "experimental" estimate can be viewed as a "leading term" in an extrapolation for a parameter of interest in an ex ante evaluation.

John DiNardo
Ford School of Public Policy
5238 Weill Hall
University of Michigan
Ann Arbor, MI 48109-3091
and NBER
jdinardo@umich.edu

David S. Lee
Industrial Relations Section
Princeton University
Firestone Library A-16-J
Princeton, NJ 08544
and NBER
davidlee@princeton.edu

1 Introduction

This chapter provides a selective review of some contemporary approaches to program evaluation. Our review is primarily motivated by the recent emergence and increasing use of the a particular kind of "program" in applied microeconomic research, the so-called Regression Discontinuity (RD) Design of Thistlethwaite and Campbell (1960). In a recent survey, Lee and Lemieux (2009) point out that the RD design has found good use in a wide variety of contexts, and that over the past decade, the way in which researchers view the approach has evolved to a point where it is now considered to yield highly credible and transparent causal inferences. At the time of the last volumes of the Handbook of Labor Economics, the RD design was viewed simultaneously as a "special case" of Instrumental Variables (IV) (Angrist and Krueger, 1999) and a "special case" of a "selection on observables", or matching approach Heckman et al. (1998b). Recent theoretical analyses and the way in which practitioners interpret RD designs reveal a different view; Lee and Lemieux (2009) point out that the RD design can be viewed as a close "cousin" of the randomized experiment. In this chapter, we provide an extended discussion of this view, and also discuss some of the issues that arise in the practical implementation of the RD design. The view of the RD design as a "cousin" of the randomized experiment leads to our second, broader objective in this review: to chart out this perspective's implicit "family tree" of commonly used program evaluation approaches.[1]

Our discussion necessarily involves a discussion of "heterogeneous treatment effects", which is one of the central issues in a wider debate about the relative merits of "structural" versus "design-based"/ "experimentalist" approaches.[2] In setting forth a particular family tree, we make no attempt to make explicit or implicit judgments about what is a "better" or "more informative" approach to conducting research. Instead, we make two distinctions that we think are helpful in our review.

First, we make a clear distinction between two very different kinds of evaluation problems. One is what could be called the *ex-post evaluation problem*, where the main goal is to document "what happened" when a particular program was implemented. The problem begins with an explicit understanding that a very particular program was run, individuals were assigned to, or self-selected into, program status in a very particular way (and we as researchers may or may not know very much about the process), and that because of the way the program was implemented, it may only be possible to identify effects for certain

[1] Other recent reviews of common evaluation approaches include, for example, Heckman and Vytlacil (2007a,b); Abbring and Heckman (2007).

[2] A sampling of papers that reflects this debate would include Heckman and Vytlacil (2005), Heckman et al. (2006), Deaton (2008), Imbens (2009), Keane (2009) and Angrist and Pischke (2010).

sub-populations. In this sense, the data and the context (the particular program) define and set limits on the causal inferences that are possible. Achieving a high degree of internal validity (a high degree of confidence that what is measured indeed represents a causal phenomenon) is the primary goal of the ex post evaluation problem.

The other evaluation problem is the *ex-ante evaluation problem,* which begins with an explicit understanding that the program that was actually run may not be the one that corresponds to a particular policy of interest. Here, the goal is not descriptive, but is instead predictive. What would be the impact if we expanded eligibility of the program? What would the effects of a similar program if it were run at a national (as opposed to a local) level? Or if it were run today (as opposed to 20 years ago)? It is essentially a problem of forecasting or extrapolating, with the goal of achieving a high degree of external validity.[3]

We recognize that in reality, no researcher will only pursue (explicitly or implicitly) one of these goals to the exclusion of the other. After all, presumably we are interested in studying the effects of a particular program that occurred in the past because we think it has predictive value for policy decisions in the here and now. Likewise, a forecasting exercise usually begins with some assessment of how well methods perform "in-sample". Nevertheless, keeping the "intermediate goals" separate allows us to discuss more clearly how to achieve those goals, without having to discuss which of them is "more important" or ambitious, or more worthy of a researcher's attention.

The second distinction we make – and one that can be more helpful than one between "structural and "design-based" approaches – is the one between "structural" and "design-based" *statistical conditions*. When we have some institutional knowledge about the process by which treatment was assigned, and when there can be common agreement about how to represent that knowledge as a statistical statement, we will label that a "D"-condition; "D" for "data-based", "design-driven", or "descriptive". These conditions are better thought of as *descriptions* of what actually generated the data, rather than *assumptions*. By contrast, when important features of the data generating process are unknown, we will have to invoke some conjectures about behavior (perhaps motivated by a particular economic model), or other aspects about the environment. When we do not literally know if the conditions actually hold, but nevertheless need them to make inferences, we will label them "S"-conditions; "S" for "structural", "subjective", or "speculative". As we shall see, inference about program effects will frequently involve a combination of "D" and "S" condi-

[3] In our chapter, we will say nothing about another kind of ex ante evaluation question: what would be the effects of a program that was never run in the first place, or of a qualitatively different kind of program? See the discussion inTodd and Wolpin (2006).

tions: it is useful to be able to distinguish between conditions whose validity is secure and those conditions whose validity is not secure.

Note that although we may not know whether "S"-conditions are literally true, sometimes they will generate strong testable implications, and sometimes they will not. And even if there is a strong link between what we know about program assignment and a "D" condition, a skeptic may prefer to treat those conditions as hypotheses; so we will also consider the testable implications that various "D"-conditions generate.

Using these distinctions, we examine some commonly employed evaluation strategies, and assess them against a common set of criteria for "internal validity". We also provide a few concrete illustrations of how the goal of an ex post evaluation are quite complementary to the that an ex ante evaluation. Specifically, for a number of the designs, where "external validity" is an issue, we show some examples where internally valid causal estimates – supplemented with specific "S"-conditions – can be viewed as a "leading term" in an extrapolation for a parameter of interest from an ex ante evaluation standpoint.

Our review of commonly employed evaluation strategies will highlight and emphasize the following ideas, some of which have long been known and understood, others that have gained much attention in the recent literature, and others that have been known for some time but perhaps have been under-appreciated:

- From an ex post evaluation standpoint, a carefully planned experiment using random assignment of program status represents the ideal scenario, delivering highly credible causal inferences. But from an ex ante evaluation standpoint, the causal inferences from a randomized experiment may be a poor forecast of what were to happen if the program where to be "scaled up". We provide a simple illustration of how this policy parameter of interest might be linked to the parameter identified from an experiment.

- When program status is described as random assignment with imperfect (and non-random) compliance, the IV (i.e. Wald) estimand delivers an average causal effect that may well not be as "local" as the usual Local Average Treatment Effect (LATE) interpretation suggests. Although LATE has been defined as the "average treatment effect for [only for] individuals whose treatment status is influenced by changing an exogenous regressor" Angrist (2004), we show that a "probabilistic monotonicity" condition allows the IV estimand to be interpreted as a weighted average effect for all individuals, where the weights are proportional to the effect of the instrument on the *probability* of treatment receipt.

- From an ex post evaluation standpoint, when program status is characterized as random assignment with imperfect compliance, LATE represents "what is possible" to identify with minimal assumptions. But from an ex ante evaluation standpoint, it may not be adequate for predicting, for example, the impact of the program if receipt was mandated (the Average Treatment Effect (ATE)). We highlight the well-known fact that ATE can be viewed as an extrapolation that has LATE as its "leading term".

- Curiously, our literature search revealed that applied researchers typically do not conduct or report such extrapolations, even though the parameters of that extrapolation are identified from the same data used to compute estimates of LATE. We apply such an extrapolation for a small sampling of studies in the literature to show the differences between LATE and (one estimate of) ATE in practice.

- The presence of "local random assignment" around the threshold in a Regression Discontinuity design is not merely a "maintained" assumption, but rather a *consequence* of a structural assumption (with strong testable implications) about the extent to which agents can precisely manipulate the assignment variable.

- The discontinuity in the RD estimand generally has a less "local" interpretation than "the average effect for those individuals at the threshold". It can be viewed as a weighted average effect, where the weights are proportional to the ex ante likelihood that the value of the individual's assignment variable would lie in a neighborhood of the threshold.

- It is clear that *any* program evaluation method ultimately requires unobservables to be independent with either an instrument or treatment status itself. But there is an important difference between *assuming* that unobservables are independent of instruments or program status, and when such a condition holds as a *consequence* of a particular data generating process.

- When employing matching estimators in a "selection on observables" approach in non-experimental settings, "adding more controls" in the analysis carries a great risk of *exacerbating* any possible selection biases.

The chapter is organized as follows: in section 2 we provide some background for our review, including our criteria for assessing various research designs; we also make some important distinctions between types of "program evaluation" that will be useful in what follows. One important distinction will be between research

designs where the investigator has detailed institutional knowledge of the process by which individuals were assigned to treatment ("dominated by knowledge of the assignment process") and those research designs where such information is lacking – what we describe as being "dominated by self–selection." In section 3, we discuss the former: this includes both randomized controlled trials and the regression discontinuity design. In section 4, we discuss the latter: this includes "differences–in–differences", instrumental variables ("selection on unobservables"), matching estimators, ("selection on observables") Section 5 concludes.

2 Scope and Background

The term "program evaluation" can be used to denote any "evaluation" of how a "program" operated. While this might include, for example, narrative descriptions by people who participated in the program, our focus will be solely on statistical and econometric evaluation. For our purposes, a program is a set of interventions, actions or "treatments" (typically binary), which are assigned to participants and are suspected of having some consequences on the outcomes experienced by the participants. Individuals who are "assigned" or "exposed" to treatment may or may not take up the treatment; when some individuals are assigned to, but do not take up the treatment we will often find it convenient to evaluate the effect of the offer of treatment (an "intent to treat analysis"), rather than the effect of the treatment *per se*, although we will examine what inferences can be made about the effect of the treatment in these situations. The problem will be to study the causal effect of the treatment when "the effects under investigation tend to be masked by fluctuations outside the experimenter's control"(Cox, 1958). Examples of programs and treatments include not only explicit social experiments such as those involving the provision of job training to individuals under the Job Training Partnership Act (JTPA) (Guttman, 1983), but also "treatments" provided outside the context of specifically designed social experiments. Some examples of the latter include the provision of collective bargaining rights to workers at firms (DiNardo and Lee, 2004), the effects of social insurance on labor market outcomes (Lemieux and Milligan, 2008), health insurance (Card et al., 2009b,a) and schooling to mothers (McCrary and Royer, 2010).

Our review will be selective. In particular, we will focus most of our attention on situations in which "institutional knowledge of the data generation process" strongly informs the statistical and econometric analysis.[4] With such a focus, a discussion of randomized controlled trials (RCTs) and the regression discon-

[4] For a comprehensive discussion and review of many of these issues see the reviews of Heckman and Vytlacil (2007a,b); Abbring and Heckman (2007).

tinuity design (RDD) are featured not because they are "best" in some single index ranking of "relevance", but because they often provide situations where a "tight link" between the posited statistical model and the institutional details of the experiment lends credibly to the conclusions. The statistical model employed to analyze a simple, well–designed RCT often bears a tighter resemblance to the institutional details of the designed experiment than does, for example, a Mincerian wage regression. In this latter case, the credibility of the exercise does not rest on the fact that wages are set in the market place as a linear combination of a non-stochastic relationship between potential experience, schooling, etc. and a stochastic error term: the credibility of such an exercise instead rests on factors *other* than its close resemblance to the institutional realities of wage setting.

The distinction between these situations has sometimes been blurred: the Neyman–Holland–Rubin Model (Splawa-Neyman et al., 1990, 1935; Rubin, 1990, 1974, 1986; Holland, 1986), which we discuss later, has been used in situations both where the investigator *does* have detailed institutional knowledge of the data generating process and where the investigator *does not*. Our focus is on "the experiment that happened" rather than the "experiment we would most like to have been conducted". As others have noted, this focus can be limiting, and a given experiment may provide only limited information (if any) on structural parameters interesting to some economists (see for example Heckman and Vytlacil (2007a)). If a designed experiment assigns a package of both "remedial education" and "job search assistance" to treated individuals, for example, we may not be able to disentangle the separate effects of each component on subsequent employment outcomes. We may be able to do better if the experiment provides random assignment of each of the components separately and together, but this will depend crucially on the experiment that was actually conducted.

In adopting such a focus, we do not mean to suggest that the types of research designs we discuss should be the only ones pursued by economists and we wish to take no position on where the "marginal research dollar" should be spent or the appropriate amount of energy which should be dedicated to "structural analyses"; for some examples of some recent contributions to this debate see Deaton (2008); Heckman and Urzua (2009); Imbens (2009); Keane (2009); Rust (2009). Moreover, even with this narrow focus there are several important subjects we will not cover, such as those involving a continuously distributed randomized instrument as in Heckman and Vytlacil (2001a); some of these issues are treated in Taber and French (2010).

2.1 Different Goals of Program Evaluation– A Broad Brush Comparison

It will be useful to reiterate a distinction that has been made elsewhere (see for example, Todd and Wolpin (2006) and Wolpin (2007)), between ex ante evaluation and ex post evaluation. *Ex post* policy evaluation occurs upon or after a policy has been implemented; information is collected about the outcomes experienced by those who participated in the "experiment" and an attempt is made to make inferences about the role of a treatment in influencing the outcomes. An ex post evaluation generally proceeds by selecting a statistical model with a tight fit to the experiment that actually happened (whether or not the experiment was "planned"). The claims that are licensed from such evaluations are context dependent – an experiment conducted among a specific group of individuals, at a specific time and specific place, may or may not be a reliable indicator of what a treatment would do among a different group of individuals at a different time or place. The credibility of an ex post evaluation depends on the credibility of the statistical model of the *experiment.* Drug trials and social experiments are examples of "planned" experiments; similarly, regression discontinuity designs, although not necessarily planned, can also often provide opportunities for an ex post evaluation.

Ex ante evaluation*,* by contrast, does not require an experiment to have happened. It is the attempt to "study the effects of policy changes prior to their implementation"(Todd and Wolpin, 2006).[5] Unlike the ex post evaluation, the credibility of an ex ante evaluation depends on the credibility of the statistical model of the *behavior* of individuals and the *environment* to which the individuals are subjected. An influential ex ante evaluation was McFadden et al. (1977), which built a random utility model to forecast the demand for the San Francisco BART subway system before it was built. In that case, the random utility model is a more or less "complete", albeit highly stylized, description of utility maximizing agents, their "preferences", etc. In short, the statistical model explains *why* individuals make their observed choices. The model of behavior and the environment *is* the data generation process.

This contrasts sharply with ex post evaluation, where apart from the description of the treatment assignment mechanism, one is as agnostic as possible about what specific behavioral model is responsible for the observed data other than the assignment mechanism. We describe this below as "pan-theoretic" – the goal in

[5] "Structural models" more generally refer to a collection of stylized mathematical descriptions of behavior and the environment which are combined to produce predictions about the effects of different choices, etc. It is a very broad area, and we make no attempt to review this literature. For a tiny sample of some of the methodological discussion, see Haavelmo (1944), Marschak (1953), Lucas (1976), Ashenfelter and Card (1982), Heckman (1991), Heckman (2000), Reiss and Wolak (2007), Heckman and Vytlacil (2007a), Deaton (2008), Fernández-Villaverde (2009), Heckman and Urzua (2009), and Keane (2009). We also ignore other types of structural models including "agent based" models (Windrum et al., 2007; Tesfatsion, 2007).

an ex post evaluation is to write down a statistical model of the assignment process or the experiment that is consistent with as broad a class of potential models as possible. When the analyst has detailed institutional knowledge of the assignment mechanism, there is usually very little discretion in the choice of statistical model – it is dictated by the the institutional details of the actual experiment. As observed by Wolpin (2007), however, this is not the case in the ex ante evaluation: "Researchers, beginning with the same question and using the same data, will generally differ along many dimensions in the modeling assumptions they make, and resulting models will tend to be indistinguishable in terms of model fit."

Since human behavior is so complicated and poorly understood (relative to the properties of simple treatment assignment mechanisms), ex ante evaluations typically place a high premium on some form of "parsimony" – some potential empirical pathways are necessarily omitted from the model. Researchers in different fields, or different economists, may construct models of the same outcomes which are very different. Because many different models – with different implications, but roughly the same "fit" to the data– might be used in an ex ante evaluation, there are a wide variety of ways in which such models are validated (See Heckman (2000); Keane and Wolpin (2007); Keane (2009) and the references therein for useful discussion). Given the goal of providing a good model of what might happen in contexts *different* than those in which the data was collected, testing or validating the model is considerably more difficult. Indeed, "the examination of models' predictive ability is not especially common in the microeconometrics literature"(Fang et al., 2007). Part of the difficulty is that by necessity, some variables in the model are "exogenous" (determined outside the model), and if these variables affect the outcome being studied, it is not sufficient to know the structure. For the ex ante evaluation to be reliable, "it is also necessary to know past and future values of all exogenous variables" (Marschak, 1953) . Finally, it is worth noting that an ex ante *evaluation* (as opposed to a mere forecasting exercise) generally requires a specification of "values" (a clear discussion of the many issues involved can be found in Heckman and Smith (1998)).

In the following table, we outline some of the similarities and differences between the two kinds of evaluations, acknowledging the difficulties of "painting with a broad brush":

Ex Post Program Evaluation	Ex Ante Program Evaluation
What did the program do? Retrospective: what happened?	What do we think a program will do? Prospective/predictive: what would happen?
Focus on the program at hand	Focus on forecasting effects of different program
For what population *do* we identify causal effect?	For what population do we *want* to identify causal effect?
Desirable to have causal inferences not reliant on specific structural framework/model	Question ill-posed without structural framework/paradigm
No value judgments on "importance" of causal facts	Some facts will be more helpful than others
Inferences require assumptions	Predictions require assumptions
Desirable to test assumptions whenever possible	Desirable to test assumptions whenever possible
Ex Ante problem guides what programs to design/analyze	Would like predictions consistent with results of Ex Post evaluation
Inference most appropriate for situations that "resemble" the experiment and are similar to that which produce the observed data	Inferences intended for situations that are different than that which produced the observed data

2.2 The Challenges of the Ex Post (Descriptive) Evaluation Problem

Here we describe a prototypical ex post program evaluation, where the perspective is that an event has occurred (i.e. some individuals were exposed to the program, while others were not) and data has been collected. The ex post evaluation question is: Given the particular program that was implemented, and the data that was collected, what is the causal effect of the program on a specific outcome of interest?

For example, suppose a state agency implements a new program that requires unemployment insurance claimants to be contacted via telephone by a job search counselor for information and advice about re-employment opportunities, and data is collected on the labor market behavior of the claimants before and after being exposed to this program. The ex post evaluation problem is to assess the impact of this particular job search program on labor market outcomes (e.g. unemployment durations) for the population of individuals to whom it was exposed.

One might also want to know what the program's impact *would be* in a different state, or 5 years from now, or for a different population (e.g. recent high school graduates, rather than the recently unemployed), or if the job counselor were to make a personal visit to the UI claimant (rather than a phone call). But in our hypothetical example none of these things happened. We consider these questions to be the concern of an *ex ante* program evaluation – a forecast of the effect of a program that *has not* occurred. For now, we consider the program that was *actually implemented*, and its effect on the population to which the program was *actually* exposed, and focus on the goal of making as credible and precise causal inferences as possible

(See Heckman and Vytlacil (2007a,b); Abbring and Heckman (2007); Keane and Wolpin (2007); Todd and Wolpin (2006) for discussion.)

We describe the general evaluation problem using the following notation:

- Y is the outcome of interest.
- D is the program, or treatment, status variable, equal to 1 if "treated" and 0 if not.
- W is a vector of all variables that could impact Y – some observable and others unobservable to the researcher – realized *prior* to the determination of program status. For example, W can represent immutable characteristics (e.g. race), constraints faced by, actions taken by, or information known to the individual, for example.
- U is a fundamentally unobservable random variable that denotes an individual's "type". By "type" we simply mean that at the time of the determination of program status, those with the same value of U can be viewed as identical agents in the sense that for those with the same U, 1) they have the same W and 2) the impact of W and D on Y is the same. U can be thought of as the equivalent to the subscript i in microeconometric analysis. In our exposition, we are instead using U so that we can consider a distribution of types. $F_U(u)$ is the cdf of U.

A general framework for the evaluation problem can be given by the system:

$$W \equiv w(U) \tag{1}$$

$$P^* \equiv p^*(W,U) \equiv \Pr[D=1|W,U] \tag{2}$$

$$Y \equiv y(D,W,U) \tag{3}$$

In the first equation, W is a random vector because U denotes the type of a randomly chosen individual from the population. With $w(\cdot)$ being a real-valued function, those with the same U (identical agents) will have the same W, but there may be variation in U conditional on an observed value of W. Furthermore, since W is determined before D, D does not enter the function $w(\cdot)$.

The second equation defines the *latent* propensity to be treated, P^*. Program status can be influenced by type U or the factors W. Additionally, by allowing P^* to take values between 0 and 1, we are allowing

for the possibility of "other factors" outside of W and U that could have impacted program status. If there are no "other factors", then P^* takes on the values 0 or 1. Since W is a function of U, U is sufficient to determine this propensity. But we include W in the function $p^*(\cdot,\cdot)$ to emphasize that one could conceive of a different action (an element of W) that could have been taken, that would have influenced P^*. W might include years of education obtained prior to exposure to the job search assistance program, and one could believe that education could impact the propensity to be a program participant. It is important to note that P^* is quite distinct from the well-known "propensity score", as we will discuss in Section 4.3. Not only is P^* potentially a function of some unobservable elements of W, but even conditional on W, P^* can vary across individuals.

The final equation is the outcome equation, with the interest centering on the impact of D on Y, keeping all other things constant. As before, even though U is sufficient to determine Y, we emphasize W (e.g. pre-program education levels) as a separate argument in the function $y(\cdot,\cdot,\cdot)$, because for the same U we can imagine Y responding to changes in W. Also, one might be interested in program effects for individuals with different values of W (e.g. race, gender).

Note that this notation has a direct correspondence to the familiar "potential outcomes framework" (Splawa-Neyman et al. (1990); Rubin (1974); Holland (1986)).[6] The framework also accommodates standard latent variable threshold-crossing models (Heckman, 1974, 1976, 1978) such as:

$$\begin{aligned} Y &= \alpha + D\beta + X\gamma + \varepsilon \\ D &= 1[X\delta + V > 0] \end{aligned}$$

where X, ε, (with an arbitrary joint distribution) are elements of W , and $P^* = \Pr[V > -X\delta | X, \varepsilon]$. The framework also corresponds to that presented in Heckman and Vytlacil (2005).[7] The key difference is that we will not presume the existence of a continuously distributed instrument Z that is independent of all the unobservables in W.

Throughout this chapter, we maintain a standard assumption in the evaluation literature (and in much of micro-econometrics) that each individual's behaviors or outcomes do not directly impact the behaviors of others (i.e., we abstract from "peer effects", general equilibrium concerns, etc.).

[6] Y_1 and Y_0 (in the potential outcomes framework) correspond to $y(1, w(U), U)$ and $y(0, w(U), U)$.

[7] Specifically, where we consider their X, U_1, and U_0 as elements of our vector W.

Define the causal effect for an individual with $U = u$ and $W = w$ as

$$\Delta(w,u) \equiv y(1,w,u) - y(0,w,u)$$

If U and all the elements of W were observed, then the causal effect could be identified at any value of W and U provided there existed some treated and non-treated individuals.

The main challenge, of course, is that the econometrician will never observe U (even if individuals can be partially distinguished through the observable elements of W). Thus, even conditional on $W = w$, it is in general *only* possible to learn something about the *distribution* of $\Delta(w,U)$. Throughout this chapter we will focus on – as does much of the evaluation literature – *average* effects

$$\int \Delta(w(u),u)\,\psi(u)\,dF_U(u) \tag{4}$$

where $\psi(u)$ is some weighting function such that $\int \psi(u)\,dF_U(u) = 1$. (See Heckman and Vytlacil (2007a); Abbring and Heckman (2007) for a discussion of distributional effects and effects other than the average.)

The source of the causal inference problem stems from unobserved heterogeneity in P^*, which will cause treated and untreated populations to be noncomparable. The treated will tend to have higher P^* (and hence the U and W that lead to high P^*), while the untreated will have lower P^* (and hence values of U and W that lead to low P^*). Since U and W determine Y, the average Y will generally be different for different populations.

More formally, we have

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \int E[y(1,w(U),U)|D=1,P^*=p^*]\,f_{P^*|D=1}(p^*)\,dp^* \\ &\quad - \int E[y(0,w(U),U)|D=0,P^*=p^*]\,f_{P^*|D=0}(p^*)\,dp^* \\ &= \int E[y(1,w(U),U)|P^*=p^*]\,f_{P^*|D=1}(p^*)\,dp^* \\ &\quad - \int E[y(0,w(U),U)|P^*=p^*]\,f_{P^*|D=0}(p^*)\,dp^* \end{aligned} \tag{5}$$

where the $f_{P^*|D=d}(p^*)$ is the density of P^* conditional on $D = d$, and the second equality follows from the fact that $E[y(d,w(U),U)|D=d,P^*=p^*] = E[y(d,w(U),U)|P^*=p^*]$: for all observations with an identical probability of receiving treatment, the distribution of unobservables will be identical between $D =$

1 and $D = 0$ populations.[8] Importantly, any nontrivial marginal density $f_{P^*}(p^*)$ will necessarily lead to $f_{P^*|D=1}(p^*) \neq f_{P^*|D=0}(p^*)$.[9]

In our discussion below, we will point out how various research designs grapple with the problem of unobserved heterogeneity in P^*. In summary, in an ex post evaluation problem, the task is to translate whatever knowledge we have about the assignment mechanism into restrictions on the functions given in Equations (1), (2), or (3), and to investigate, as a result, what causal effects can be identified from the data.

2.2.1 Criteria for Internal Validity and the Role of Economic Theory

We argue that in an ex post evaluation of a program, the goal is to make causal inferences with a high degree of "internal validity": the aim is to make credible inferences and qualify them as precisely as possible. In such a descriptive exercise, the degree of "external validity" is irrelevant. On the other hand, "external validity" will be of paramount importance when one wants to make predictive statements about the impact of the same program on a different population, or when one wants to use the inferences to make guesses about the possible effects of a slightly different program. That is, we view "external validity" to be the central issue in an attempt to use the results of an ex post evaluation for an ex ante program evaluation; we further discuss this in the next section.

What constitutes an inference with high "internal validity"?[10] Throughout this chapter we will consider three criteria. The first is the extent to which there is a tight correspondence between what we know about the assignment-to-treatment mechanism and our statistical model of the process. In some cases, the assignment mechanism might leave very little room as to how it is to be formally translated into a statistical assumption. In other cases, little might be known about the process leading to treatment status, leaving much more discretion in the hands of the analyst to model the process. We view this discretion as potentially expanding the set of "plausible" (yet different) inferences that can be made, and hence generating doubt as to which one is correct.

The second criterion is the broadness of the class of models with which the causal inferences are consistent. Ideally, one would like to make a causal inference that is consistent with any conceivable behavioral model. By this criterion, it would be undesirable to make a causal inference that is only valid if a very spe-

[8] Formally, $F_{U|D=1,P^*=p^*}(u) = \frac{\Pr[D=1|U\leq u,P^*=p^*]F_{U|P^*=p^*}(u)}{\Pr[D=1|P^*=p^*]} = F_{U|P^*=p^*}(u)$, and similarly, $F_{U|D=0,P^*=p^*}(u) = F_{U|P^*=p^*}(u)$.

[9] From Bayes' rule we have $f_{P^*|D=1}(p^*) = \frac{\Pr[D=1|P^*=p^*]f_{P^*}(p^*)}{\Pr[D=1]} = \frac{p^* f_{P^*}(p^*)}{\Pr[D=1]}$, and $f_{P^*|D=0}(p^*) = \frac{(1-p^*)f_{P^*}(p^*)}{1-\Pr[D=1]}$.

[10] Campbell and Cook (1979) contains a discussion of various "threats" to internal validity.

cific behavioral model is true, and it is unknown how the inferences would change under plausible deviations from the model in question.

The last criterion we will consider is the extent to which the research design is testable; that is, the extent to which we can treat the proposed treatment assignment mechanism as a null hypothesis that could, in principle, be falsified with data (e.g. probabilistically, via a formal statistical test).

Overall, if one were to adopt these three criteria, then a research design would have low "internal validity" when 1) the statistical model is not based on what is actually known about the treatment assignment mechanism, but based entirely on speculation, 2) inferences are known only to be valid for one specific behavioral model amongst many other plausible alternatives and 3) there is no way to test the key assumption that achieves identification.

What is the role of economic (for that matter, any other) theory in the ex post evaluation problem? First of all, economic theories motivate what outcomes we wish to examine, and what causal relationships we wish to explore. For example, our models of job search (see McCall and McCall (2008) for example) may motivate us to examine the impact of a change in benefit levels on unemployment duration. Or if we were interested in the likely impacts of the "program" of a hike in the minimum wage, economists are likely to be most interested in the impact on employment, either for the purposes of measuring demand elasticities, or perhaps assessing the empirical relevance of a perfectly competitive labor market against that of a market in which firms face upward-sloping labor supply curves (Card and Krueger, 1995; Manning, 2003).

Second, when our institutional knowledge does not put enough structure on the problem to identify any causal effects, then assumptions about individuals' behavior *must* be made to make any causal statement, however conditional and qualified. In this way, structural assumptions motivated by economic theory can help "fill in the gaps" in the knowledge of the treatment assignment process.

Overall, in an ex post evaluation, the imposition of structural assumptions motivated by economic theory is done out of necessity. The ideal is to conjecture *as little as possible* about individuals' behavior so as to make the causal inferences valid under the broadest class of all possible models. For example, one could imagine beginning with a simple Rosen-type model of schooling with wealth maximization (Rosen, 1987) as a basis for empirically estimating the impact of a college subsidy program on educational attainment and lifetime earnings. The problem with such an approach is that this would raise the question as to whether the causal inferences entirely depend on that particular Rosen-type model. What if one added consumption decisions to the model? What about saving and borrowing? What if there are credit constraints? What if

there are unpredictable shocks to non-labor income? What if agents maximize present discounted utility rather than discounted lifetime wealth? The possible permutations go on and on.

It is tempting to reason that we have no choice but to adopt a specific model of economic behavior and to admit that causal inferences are conditional only on the model being true; that the only alternative is to make causal inferences that depend on assumptions that we do not even know we are making.[11] But this reasoning equates the *specificity* of a model with its *completeness*, which we believe to be very different notions.

Suppose, for example – in the context of evaluating the impact of our hypothetical job search assistance program – that the type of a randomly drawn individual from the population is given by the random variable U (with a cdf $F_U(u)$), that W represents all the constraints and actions the individual takes prior to, and in anticipation of, the determination of participating in the program D, and that outcomes are determined by the system given by Equations (1), (2), and (3). While there is no discussion of utility functions, production functions, information sets, or discount rates, the fact is that this is a *complete* model of the data generating process; that is, we have enough information to derive expressions for the joint distribution of the observables (Y,D,W) from the primitives of $F_U(u)$ and (1), (2), and (3). At the same time it is not a very *specific* (or economic) model, but in fact, quite the opposite: it is perhaps the most general formulation that one could consider. It is difficult to imagine any economic model – including a standard job search model – being inconsistent with this framework.

Another example of this can be seen in the context of the impact of a job training program on earnings. One of the many different economic structures consistent with (1), (2), and (3) is a Roy-type model of self-selection (Roy, 1951; Heckman and Honore, 1990) into training.[12] The Roy-type model is certainly *specific*, assuming perfect foresight on earnings in both the "training" or "no-training" regimes, as well as income maximization behavior. If one obtains causal inferences in the Roy model framework, an open question would be how the inferences change under different theoretical frameworks (e.g. a job search-type model, where training shifts the wage offer distribution upward). But if we can show that the causal inferences are valid within the more general – but nonetheless *complete* – formulation of (1), (2), and (3), then we know the inferences will still hold under both the Roy-type model, a job search model, or any number of plausible alternative economic theories.

[11] See Keane (2009); Rosenzweig and Wolpin (2000) for a discussion along these lines.

[12] W could be observable components of human capital, $p^*(w,u) = 1[y(1,w(u),u) - y(0,w(u),u) \geq 0]$.

2.3 The "Parameter of Interest" in an Ex Ante (Predictive) Evaluation Problem

We now consider a particular kind of predictive, or ex ante, evaluation problem: suppose the researcher is interested in predicting the effects of a program "out of sample". For example, the impact of the Job Corps Training program on the earnings of youth in 1983 in the 10 largest metropolitan areas in the U.S. may be the focus of an ex post evaluation, simply because the data at hand comes from such a setting. But it is natural to ask any one or a combination of the following questions: What would be the impact today (or some date in the future)? What would be the impact of an expanded version of the program in more cities (as opposed to the limited number of sites in the data)? What would be the impact on an older group of participants (as opposed to only the youth)? What would be the impact of a program that expanded eligibility for the program? These are examples of the questions that are in the domain of an ex ante evaluation problem.

Note that while the ex post evaluation problem has a descriptive motivation – the above questions implicitly have a prescriptive motivation. After all, there seems no other practical reason why knowing the impact of the program "today" would be any "better" than knowing the impact of the program 20 years ago, other than because such knowledge helps us make a particular policy decision today. Similarly, the only reason we would deem it "better" to know the impact for an older group of participants, or participants from less disadvantaged backgrounds, or participants in a broader group of cities is because we would like to evaluate whether actually targeting the program along any of these dimensions would be a good idea.

One can characterize an important distinction between the ex post and ex ante evaluation problems in terms of Equation (4). In an ex post evaluation, the weights $\psi(u)$ are dictated by the constraints of the available data, and what causal effects are most plausibly identified. It is simply accepted as a fact – however disappointing it may be to the researcher – that there are only a few different weighted average effects that can be plausibly identified, whatever weights $\psi(u)$ they involve. By contrast, in an ex ante evaluation, the weights $\psi(w)$ are *chosen* by the researcher, irrespective of the feasibility of attaining the implied weighted average "of interest". These weights may reflect the researcher's subjective judgement about what is an "interesting" population to study. Alternatively, they may be implied by a specific normative framework. A clear example of the latter is found in Heckman and Vytlacil (2005), who begin with a Benthamite social welfare function to define a "policy relevant treatment effect", which is a weighted average treatment effect with a particular form for the weights $\psi(u)$.

One can thus view "external validity" to be the degree of similarity between the weights characterized

in the ex post evaluation and the weights defined as being "of interest" in an ex ante evaluation. From this perspective, any claim about whether a particular causal inference is "externally valid" is necessarily imprecise without a clear definition of the desired weights and their theoretical justification. Again, the PRTE of Heckman and Vytlacil (2005) is a nice example where such a precise justification is given.

Overall, in contrast to the ex post evaluation, the goals of an ex ante evaluation are not necessarily tied to the specific context of or data collected on any particular program. In some cases, the researcher may be interested in the likely effects of a program on a population for which the program was already implemented; the goals of the ex post and ex ante evaluation would then be similar. But in other cases, the researcher may have reason to be interested in the likely effects of the program on different populations or in different "economic environments"; in these cases ex post and ex ante evaluations – even when they use the same data – would be expected to yield different results. It should be clear that however credible or reliable the ex post causal inferences are, ex ante evaluations using the same data will necessarily be more speculative and dependent on more assumptions, just as forecasting out of sample is a more speculative exercise than within-sample prediction.

2.3.1 Using Ex Post Evaluations for Ex Ante Predictions

In this chapter, we focus most of our attention on the goals of the ex post evaluation problem, that of achieving a high degree of internal validity. We recognize that the weighted average effects that are often identified in ex post evaluation research designs may not correspond to a potentially more intuitive "parameter of interest", raising the issue of "external validity". Accordingly – using well-known results in the econometric and evaluation literature – we sketch out a few approaches for extrapolating from the average effects obtained from the ex post analysis to effects that might be the focus of an ex ante evaluation.

Throughout the chapter, we limit ourselves to contexts in which a potential instrument is binary, because the real-world examples where potential instruments have been explicitly or "naturally" randomized, the instrument is invariably binary. As is well-understood in the evaluation literature, this creates a gap between what causal effects we *can* estimate and the potentially more "general" average effects of interest. It is intuitive that such a gap would diminish if one had access to an instrumental variable that is continuously distributed. Indeed, as Heckman and Vytlacil (2005) show, when the instrument Z is essentially randomized (and excluded from the outcome equation) *and* continuously distributed in such a way that that $\Pr[D=1|Z=1]$ is continuously distributed on the unit interval, then the full set of what they define as

Marginal Treatment Effects (MTE) can be used construct various policy parameters of interest.

3 Research Designs Dominated by Knowledge of Assignment Process

In this section, we consider a group of research designs in which the model for the data generating process is to a large extent dictated by explicit institutional knowledge of how treatment status was assigned. We make the case that these four well-known cases deliver causal inferences with a high degree of "internal validity" because of at least three reasons: 1) some important or all aspects of the econometric model is a literal description of the treatment assignment process, 2) the validity of the causal inferences hold true within a seemingly broad class of competing behavioral models, and perhaps most importantly, 3) the statistical statements that describe the assignment process simultaneously generate strong observable predictions in the data. For these reasons, we argue that these cases might be considered "high-grade" experiments/natural experiments.[13]

In this section, we also consider the issue of "external validity" and the ex ante evaluation problem. It is well understood that in the four cases below, the populations for which average causal effects are identified may not correspond to the "populations of interest". The ATE identified in a small, randomized experiment does not necessarily reflect the impact of a widespread implementation of the program; the Local Average Treatment Effect (LATE) of Imbens and Angrist (1994) is distinct from the ATE; the causal effect identified by the Regression Discontinuity Design of Thistlethwaite and Campbell (1960) does not reflect the effect of making the program available to individuals whose assignment variable is well below the discontinuity threshold. For each case, we illustrate how imposing some structure on the problem can provide an explicit link between the quantities identified in the ex post evaluation and the parameters of interest in an ex ante evaluation problem.

3.1 Random Assignment with Perfect Compliance

3.1.1 Simple Random Assignment

We start by considering simple random assignment with perfect compliance. "Perfect compliance" refers to the case that individuals who are assigned a particular treatment, do indeed receive the treatment. For example, consider a re-employment program for unemployment insurance claimants, where the "program"

[13] A discussion of "low grade" experiments can be found in Keane (2009). See also Rosenzweig and Wolpin (2000).

is being contacted (via telephone and/or personal visit) by a career counselor, who provides information that facilitates the job search process. Here, participation in this "program" is not voluntary, and it is easy to imagine a public agency randomly choosing a subset of the population of UI claimants to receive this treatment. The outcome might be time to re-employment or total earnings in a period following the treatment.

In terms of the framework defined by equations (1), (2), and (3), this situation can be formally represented as

- D1: (Simple Random Assignment): $P^* = p_0$, $p_0 \in (0,1)$, a nonrandom constant

That is, for the entire population being studied, every individual has the same probability of being assigned to the program.

It is immediately clear that the distribution of P^* becomes degenerate, with a single mass point at $P^* = p_0$, and so the difference in the means in Equation (5) becomes

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[y(1,w(U),U)] - E[y(0,w(U),U)] \\ &= E[\Delta(w(U),U)] \\ &= \int \Delta(w(u),u)\, dF_U(u) \equiv ATE \end{aligned}$$

where the ATE is the "average treatment effect". The weights from Equation (4) are $\psi(u) = 1$ in this case. A key problem posed in Equation (5) is the potential relationship between the latent propensity P^* and Y (the functions $E[y(1,w(U),U)|P^* = p^*]$ and $E[y(0,w(U),U)|P^* = p^*]$). Pure random assignment "solves" the problem by *eliminating* all variation in P^*.

Internal Validity: Pan-theoretic Causal Inference

Let us now assess this research design on the basis of the three criteria described in Section 2.2.1. First, given the general formulation of the problem in Equations (1), (2), and (3), Condition D1 is much less an assumption, but rather a literal description of the assignment process – the "D" denotes a *descriptive* element of the data generating process. Indeed, it is not clear how else one would formally describe the randomized experiment.

Second, the causal inference is apparently valid for any model that is consistent with the structure given in Equations (1), (2), and (3). As discussed in Section 2.2.1, it is difficult to conceive of a model of behavior that would *not* be consistent with (1), (2), and (3). So even though we are not explicitly laying out

the elements of a specific model of behavior (e.g. a job search model), it should be clear that given the distribution $F_U(U)$, Equations (1), (2), and (3), and Condition D1 constitutes a *complete* model of the data generating process, and that causal inference is far from being "atheoretic". Indeed, the causal inference is best described as "pan-theoretic", consistent with a broad – arguably broadest – class of possible behavioral models.

Finally, and perhaps most crucially, even though one could consider D1 to be a *descriptive* statement, we could alternatively treat it as a *hypothesis*, one with testable implications. Specifically, D1 implies

$$\begin{aligned} F_{U|D=1}(u) &= \frac{\Pr[D=1|U\leq u]F_U(u)}{\Pr[D=1]} \\ &= \frac{p_0 F_U(u)}{\Pr[D=1]} = F_U(u) \end{aligned} \tag{6}$$

and a similarly, $F_{U|D=0}(u) = F_U(u)$. That is, the distribution of unobserved "types" is identical in the treatment and control groups. Since U is unobservable, this itself is not testable. But a direct consequence of result is that the pre-determined characteristics/actions must be identical between the two groups as well,

$$\begin{aligned} F_{W|D=d}(w) &= \Pr[w(U)\leq w|D=d] \\ &= \Pr[w(U)\leq w] \end{aligned}$$

which *is* a testable implication (as long as there are some observable elements of W).

The implication that the entire joint distribution of *all* pre-determined characteristics be identical in both the treatment and control states is indeed quite a stringent test, and also independent of any model of the determination of W. It is difficult to imagine a more stringent test.

Although it may be tempting to conclude that "even random assignment must assume that the unobservables are uncorrelated with treatment", on the contrary, the key point here is that the balance of unobservable types U between the treatment and control groups is not a primitive *assumption;* instead, it is a direct *consequence* of the assignment mechanism, which is described by D1. Furthermore, balance in the observable elements of W is *not* an additional assumption, but a natural implication of balance in the unobservable type U.

One might also find D1 "unappealing" since mathematically it seems like a strong condition. But from an ex post evaluation perspective, whether D1 is a "strong" or "weak" condition is not as important as the

fact that D1 is beyond *realistic*: it is practically a literal description of the randomizing process.

3.1.2 Stratified/Block Randomization

Now, suppose there is a subset of elements in W – call this vector X – that are observed by the experimenter. A minor variant on the above mechanism is when the probability of assignment to treatment is different for different groups defined by X, but the probability of treatment is identical for all individuals *within* each group defined by X.[14] In our hypothetical job search assistance experiment, we could imagine initially stratifying the study population by their previous unemployment spell history: "short", "medium", and "long"-(predicted) spell UI claimants. This assignment procedure can be described as

- D2: (Random Assignment Conditional on X) $P^* = p^*(X)$, $p^*(x) \in (0,1)\ \forall x$

In this case, where there may be substantial variation in the unobservable type U for a given X, the probability of receiving treatment is identical for everyone with the same X.

The results from simple random assignment naturally follow,

$$\begin{aligned} E[Y|D=1,X=x] - E[Y|D=0,X=x] &= E[\Delta(W,U)|X=x] \\ &= \int \Delta(w(u),u)\, dF_{U|X=x}(u) \end{aligned}$$

, essentially an average treatment effect, conditional on $X = x$.

We mention this case not because D2 is a weaker, and hence more palatable assumption. Rather, it is useful to know that the statement in D2 – like the mechanism described by D1– is one that typically occurs when randomized experiments are implemented. For example, in the Negative Income Tax Experiments (Robins, 1985; Ashenfelter and Plant, 1990), X were the pre-experimental incomes, and families were randomized into the various treatment groups with varying probabilities, but those probabilities were identical for every unit with the same X. Another example is the Moving to Opportunity Experiment (Orr et al., 2003), which investigated the impact of individuals moving to a more economically advantaged neighborhood. The experiment was done in 5 different cities (Baltimore, Boston, Chicago, Los Angeles, and New

[14] While this setup has been described as the "selection on observables", "potential outcomes", "switching regressions" or "Neyman-Rubin-Holland model" Splawa-Neyman et al. (1990); Lehmann and Jr. (1964); Quandt (1958, 1972); Rubin (1974); Barnow et al. (1976); Holland (1986), to avoid confusion we will reserve the phrase "selection on observables" for the case where the investigator does not have detailed institutional knowledge of the selection process and treat the stratified/block randomization case as special case of simple randomization.

York) over the period 1994 - 1998. Unanticipated variation in the rate at which people found eligible leases led them to change the fraction of individuals randomly assigned to the treatments two different times during the experiment (Orr et al., 2003, page 232). In this case, families were divided into different "blocks" or "strata" by location $\times$ time and there was a different randomization ratio for each of these blocks.

This design – being very similar to the simple randomization case – would have a similar level of internal validity, according to two of our three criteria. Whether this design is testable (the third criterion we are considering) depends on the available data. By the same argument as in the simple random assignment case, we have

$$\begin{aligned} F_{W|D=d,X=x}(w) &= \Pr[w(U) \leq w | D = d, X = x] \\ &= \Pr[w(U) \leq w | X = x] \end{aligned}$$

So if the conditional randomization scheme is based on *all* of the Xs that are observed by the analyst, then there are no testable implications. On the other hand, if there are additional in elements in W that are observed (but not used in the stratification), then once again, one can treat D2 as a hypothesis, and test that hypothesis by examining whether the distribution of those extra variables are the same in the treated and control groups (conditional on X).

3.1.3 The Randomized Experiment: Pre-Specified Research Design and a Chance Setup

We have focused so far on the role that randomization (as described by D1 or D2) plays in ensuring a balance of the unobservable types in the treated and control groups, and have argued that in principle, this can deliver causal inferences with a high degree of internal validity.

Another characteristic of the randomized experiment is that it can be described as "pre-specified" research design. In principle, *before* the experiment is carried out, the researcher is able to dictate in advance what analyses are to be performed. Indeed, in medical research conducted in the U.S., prior to conducting an medical experiment, investigators will frequently post a complete description of the experiment in advance at a web site such as clincaltrials.gov. This posting includes how the randomization will be performed, the rules for selecting subjects, the outcomes that will be investigated, and what statistical tests will be performed. Among other things, such pre-announcement prevents the possibility of "selective reporting" – reporting the results only from those trials that achieve the "desired" result. The underlying notion motivating such

procedure has been described as providing a "severe test" – a test which "provides an overwhelmingly good chance of revealing the presence of a specific error, if it exists — but not otherwise" (Mayo, 1996, page 7). This notion conveys the idea that convincing statistical evidence does not rely *only* on the "fit" of the data to a particular hypothesis but on the *procedure* used to arrive at the result. Good procedures are ones that make fewer "errors."

It should be recognized, of course, that this "ideal" of pre-specification is rarely implemented in social experiments in economics. In the empirical analysis of randomized evaluations, analysts often cannot help but be interested in the effects for different sub-groups (in which they were not initially interested), and the analysis can soon resemble a data-mining exercise.[15] That said, the problem of data-mining is not specific to randomized experiments, and a researcher armed with a lot of explanatory variables in a non-experimental setting can easily find many "significant" results even among purely randomly generated "data" (see Freedman (1983) for one illustration). It is probably constructive to consider that there is a spectrum of pre-specification, with the pre-announcement procedure described above on one extreme, and specification searching and "significance hunting" with non-experimental data on the other. In our discussion below, we make the case that detailed knowledge of the assignment-to-treatment process can serve much the same role as a pre-specified research design in "planned" experiments – as a kind of "straight jacket" which largely dictates the nature of statistical analysis.

Another noteworthy consequence of this particular data generating process is that is essentially a "statistical machine" or a "chance set up" (Hacking, 1965) whose "operating characteristics" or statistical properties are well-understood, such as a coin flip. Indeed, after a randomizer assigns n individuals to the (well-defined) treatment, and n individuals to the control for a total of $N = 2n$ individuals, one can conduct a non-parametric *exact* test of sharp null hypothesis that does not require *any* particular distributional assumptions.

Consider the sharp null hypothesis that there is no treatment effect for any individuals (which implies that the two samples are drawn from the same distribution). In this case the assignment of the label "treatment" or "control" is arbitrary. In this example there are $P = \binom{2n}{n}$ different ways the labels "treatment" and "control" *could have* been assigned. Now consider the following procedure:

1. Compute the difference in means (or any other interesting test statistic). Call this $\widehat{\Delta}$.

[15] See Deaton (2008).

2. Permute the label treatment or control and compute the test statistic under this assignment of labels. This will generate P different values of the test statistic Δ_p^* for $p = 1 \ldots P$. These collection of these observations yield an exact distribution of the test statistic.

3. One can compute the p-value such that the probability that a draw from this distribution would exceed $\left|\hat{\Delta}\right|$.

This particular "randomization" or "permutation" test was originally proposed by Fisher (1935) for its utility to "supply confirmation whenever, rightly or, more often wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality." (Fisher, 1966, page 48) (See Lehmann (1959, pages 183–192) for a detailed discussion). Our purpose in introducing it here is *not* to advocate for randomization inference as an "all purpose" solution for hypothesis testing; rather our purpose is to show just how powerful detailed institutional knowledge of the DGP can be.

3.1.4 Ex Ante Evaluation: Predicting the Effects of an Expansion in the Program

Up to this point, with our focus on an ex post evaluation we have considered the question, "For the individuals exposed to the randomized evaluation, what was the impact of the program?" We now consider a particular ex ante evaluation question, "What would be the impact of a full-scale implementation of the program?", in a context when that full-scale implementation has not occurred. It is not difficult to imagine that the individuals who participate in a small-scale randomized evaluation may differ from those who would receive treatment under full-scale implementation. One could take the perspective that this therefore makes the highly credible/internally valid causal inferences from the randomized evaluation irrelevant, and hence that there is no choice but to pursue non-experimental methods, such as a structural modeling approach to evaluation, to answer the "real" question of interest.

Here we present an alternative view that this ex ante evaluation question is an extrapolation problem. And far from being irrelevant, estimates from a randomized evaluation can *form the basis* for such an extrapolation. And rather than viewing structural modeling and estimation as an alternative or substitute for experimental methods, we consider the two approaches to be potentially quite complementary in carrying out this extrapolation. That is, one can adopt certain assumptions about behavior and the structure of the economy to make precise the linkage between highly credible impact estimates from a small-scale experiment and the impact of a hypothetical full-scale implementation.

We illustrate this with the following example. Suppose one conducted a small-scale randomized evaluation of a job training program where participation in the *experimental study* was voluntary, while actual receipt of training was randomized. The question is, what would be the impact on earnings if we opened up the program so that participation in the *program* was voluntary?

First, let us define the parameter of interest as

$$E\left[Y^T\right] - E\left[Y^N\right]$$

where Y^T and Y^N are the earnings of a randomly drawn individual under two regimes: full-scale implementation of the program (T), or no program at all (N). This corresponds to the parameter of interest that motivates the Policy Relevant Treatment Effect (PRTE) of Heckman and Vytlacil (2001b). We might like to know the average earnings gain for everyone in the population. We can also express this as

$$\begin{aligned} E\left[Y^T\right] - E\left[Y^N\right] &= E\left[D^T Y_1^T + \left(1 - D^T\right) Y_0^T\right] - E\left[Y_0^N\right] \\ &= E\left[Y_1^T | D^T = 1\right] \Pr\left[D^T = 1\right] + E\left[Y_0^T | D^T = 0\right] \Pr\left[D^T = 0\right] - E\left[Y_0^N\right] \end{aligned} \tag{7}$$

where the D^T is the treatment status indicator in the T regime, and the subscripts denote the potential outcomes.

Make the following assumptions:

- S1 (Linear Production Technology): $Q = \sum_{j=1}^{K} a(j) L(j)$, where Q is the amount of output, $L(j)$ is the total amount of labor supplied by workers with j units of human capital, and $a(j)$ are technological parameters with $a(j+1) > a(j)$.

- S2 (Job Training as Human Capital): the random variable J is the individual's endowment of human capital, and Δ is the gain in human capital due to training, so that in the implementation regime, human capital is $J + \Delta D^T$.

- S3 (Inelastic Labor Supply): Each individual inelastically supplies L units of labor.

- S4 (Profit maximizing price-taking): $W = a\left(J + \Delta \cdot D^T\right)$.

This setup will imply that

$$Y_0^N \quad = \quad W \cdot L = a(J) \cdot L = Y_0^T$$

. Thus, S1 through S4 are simply a set of economic assumptions that says potential outcomes are unaffected by the implementation of the program; this corresponds to what Heckman and Vytlacil (2005) call *policy invariance*. This policy invariance comes about because of the linear production technology, which implies that wages are determined by the technological parameters, and *not* the supply of labor for each level of human capital.

With this invariance, we may suppress the superscript T for the potential outcomes; Equation (7) will become

$$\left(E\left[Y_1|D^T=1\right]-E\left[Y_0|D^T=1\right]\right)\Pr\left[D^T=1\right]=\left(E\left[Y_1-Y_0|D^T=1\right]\right)\Pr\left[D^T=1\right]$$

[16] Note that the key causal parameter $E\left[Y_1-Y_0|D^T=1\right]$ will in general be different from $E\left[Y_1-Y_0|D^E=1\right]$, where D^E is the indicator for having participated in the smaller scale randomized experiment (bearing the risk of not being selected for treatment). That is, the concern is that those who participate in the experimental study may not be representative of the population that would eventually participate in a full-scale implementation.

How could they be linked? Consider the additional assumptions

- S5 (Income Maximization; Perfect Information; Selection on Gains): $D^T=1$ iff $Y_1-Y_0>c_f+c_p$, with c_f is the "fixed" cost to applying for the program, and c_p is the monetary cost to the individual from receiving the treatment.
- S6 (Risk Neutrality) Individuals maximize expected income.

Together, S5 and S6 imply that we could characterize the selection into the program in the experimental regime as

$$D^E=1 \text{ iff } p(Y_1-Y_0)>c_f+pc_p$$

where p is the probability of being randomized into receiving the treatment (conditional on participating

[16] This is because $E\left[Y_0^N\right]=E\left[Y_0^N|D^T=1\right]\Pr\left[D^T=1\right]+E\left[Y_0^N|D^T=0\right]\Pr\left[D^T=0\right]$.

in the experimental study). Note that this presumes that in the experimental regime, all individuals in the population have the option of signing up for the experimental evaluation.

Finally, assume a functional form for the distribution of training effects in the population:

- S7 (Functional Form: Normality): $Y_1 - Y_0$ is normally distributed with variance σ^2.

Applying assumption S7 yields the following expressions

$$\begin{aligned} E\left[Y_1 - Y_0 | D^T = 1\right] &= E\left[Y_1 - Y_0\right] + \sigma \frac{\phi\left(\frac{c_f + c_p - E[Y_1 - Y_0]}{\sigma}\right)}{1 - \Phi\left(\frac{c_f + c_p - E[Y_1 - Y_0]}{\sigma}\right)} \\ E\left[Y_1 - Y_0 | D^E = 1\right] &= E\left[Y_1 - Y_0\right] + \sigma \frac{\phi\left(\frac{c_f + \frac{c_f}{p} - E[Y_1 - Y_0]}{\sigma}\right)}{1 - \Phi\left(\frac{c_f + \frac{c_f}{p} - E[Y_1 - Y_0]}{\sigma}\right)} \end{aligned} \tag{8}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf, respectively.

The probability of assignment to treatment in the experimental regime, p, characterizes the scale of the program. The smaller p is, the smaller the expected gain to participating in the experimental study, and hence the average effect of the study participants will be more positively selected. On the other hand, as p approaches 1, the experimental estimate approaches the policy parameter of interest because the experiment *becomes* the program of interest. Although we are considering the problem of predicting a "scaling up" of the program, this is an interesting case to case to consider because it implies that for *an already existing* program, one can potentially conduct a randomized evaluation, where a small fraction of individuals are denied the program (p close to 1), and the resulting experimentally identified effect $E\left[Y_1 - Y_0 | D^E = 1\right]$ can be directly used to predict the aggregate impact of completely shutting down the program.[17]

The left-hand side of the first equation is the "parameter of interest" (i.e. what we want to know) in an ex ante evaluation problem. The left-hand side of the second equation is "what can be identified" (i.e. what we do know) from the experimental data in the ex post evaluation problem. The latter may not be "economically interesting" per se, but at the same time it is far from being unrelated to the former.

Indeed, the average treatment effect identified from the randomized experiment is the starting point or

[17] Heckman and Vytlacil (2005) make this point clearly, noting that the treatment on the treated parameter is the key ingredient to predicting the impacts of shutting down the program.

"leading term", when we combine the above two expressions to yield

$$E\left[Y_1 - Y_0 | D^T = 1\right] = E\left[Y_1 - Y_0 | D^E = 1\right] + \sigma \left[\frac{\phi\left(\Phi^{-1}\left(1 - \Pr\left[D^T = 1\right]\right)\right)}{\Pr\left[D^T = 1\right]} - \frac{\phi\left(\Phi^{-1}\left(1 - \Pr\left[D^E = 1\right]\right)\right)}{\Pr\left[D^E = 1\right]} \right]$$

with the only unknown parameters in this expression being $\Pr\left[D^T = 1\right]$, the predicted take-up of in a full-scale implementation, and σ, the degree of heterogeneity of the potential training effects *in the entire population*. It is intuitive that *any* ex ante evaluation of the full-scale implementation that has not yet occurred will, at a minimum, need these two quantities.

In presenting this example, we do not mean to assert that the economic assumptions S1 through S7 are particularly realistic. Nor do we assert they are minimally sufficient to lead to an extrapolative expression. There are as many different different ways to model the economy as there are economists (and probably more!). Instead, we are simply illustrating that an ex ante evaluation attempt can directly *use* the results of an ex post evaluation, and in this way the description of the data generating process in an ex post evaluation (D1 or D2) can be quite complementary to the structural economic assumptions (S1 through S7). D1 is the key assumption that helps you identify whether there is credible evidence – arguably the most credible that is possible – of a causal phenomenon, while S1 through S7 provides a precise framework to think about making educated guesses about the effects of a program that has yet to be implemented. Although $E\left[Y_1 - Y_0 | D^E = 1\right]$ may not be of direct interest, obtaining credible estimates of this quantity would seem helpful for making a prediction about $E\left[Y_1 - Y_0 | D^T = 1\right]$.

3.2 Random Assignment: Imperfect Compliance

We now consider another data generating process that we know often occurs in reality – when there is randomization in the "intent to treat", but where participation in the program is potentially non-random and driven by self-selection. To return to our hypothetical job search assistance program, instead of mandating the treatment (personal visit/phone call from a career counselor), one could make participation in receiving such a call voluntary. Furthermore, one could take UI claimants and randomize them into two groups: one group receives information about the existence of this program, and the other does not receive the information. One can easily imagine that those who voluntarily sign up to be contacted by the job counselor might be systematically different from those who do not, and in ways related to the outcome. One can also imagine being interested in knowing the "overall effect" of "providing information about the program", but

more often it is the case that we are interested in participation in the program *per se* (the treatment of "being contacted by the job counselor").

We discuss this widely known data generating process within the very general framework described by Equations (1), (2), and (3). We will introduce a more accommodating monotonicity condition than that employed in Imbens and Angrist (1994) and Angrist et al. (1996). When we do so, the familiar "Wald" estimand will give an interpretation of an average treatment effect that is not quite as "local" as implied by the "local average treatment effect" (LATE), which is described as "the average treatment effect for the [subpopulation of] individuals whose treatment status is influenced by changing an exogenous regressor that satisfies an exclusion restriction" Imbens and Angrist (1994).

We begin with describing random assignment of the "intent to treat" as

- D3 (Random Assignment of Binary Instrument): $\Pr[Z=1|U=u]=p_{z1}\in(0,1)$, a nonrandom constant. We can thus write $P^*=p^*(W,U)=p_{z1}p_1^*(w(U),U)+(1-p_{z1})p_0^*(w(U),U)$ where $p_z^*(w(u),u)\equiv\Pr[D=1|U=u,Z=z]$.

This is analogous to D1 (and D2), except that instead of randomizing the treatment, we are randomizing the instrumental variable. Like D1 and D2, it is appropriate to consider this a *description* of the process when we know that Z has been randomized.

Since we have introduced a new variable Z, we must specify how it relates to the other variables:

- S8 (Excludability): $Y=y(D,W,U)$, $W=w(U)$ (Z is not included as an argument in either function).

Although this is a re-statement of Equations (1) and (3), given the existence of Z, this is a substantive and crucial assumption. It is the standard excludability condition: Z cannot have an impact on Y, either directly or indirectly through influencing the other factors W. It is typically *not* a literal descriptive statement in the way that D1 through D3 can sometimes be. It is a structural ("S") assumption on the same level as S1 through S7 and it may or may not be plausible depending on the context.

Finally, we have

- S9 (Probabilistic Monotonicity): $p_1^*(w(u),u)\geq p_0^*(w(u),u)$ for all u.

S9 is a generalization of the monotonicity condition used in Imbens and Angrist (1994); Angrist et al. (1996). In those papers, $p_1^*(w(u),u)$ or $p_0^*(w(u),u)$ take on the values 1 or 0; that is, for a given individual type U, their treatment status is *deterministic* for a given value of the instrument Z. This would imply that P^* would

have a distribution with three points of support: 0 (the latent propensity for "never-takers"), p_{z1} (the latent propensity for "compliers"), and 1 (the latent propensity for "always-takers").[18]

In the slightly more general framework presented here, for each type U, for a given value of the instrument Z, treatment status is allowed to be *probabilistic*: some *fraction* (potentially strictly between 0 and 1) of them will be treated. P^* can thus take on a continuum of values between 0 and 1. The probabilistic nature of the treatment assignment can be interpreted in at least two ways: 1) for a particular individual of type U, there are random shocks beyond the individual's control that introduces some uncertainty into the treatment receipt (e.g. there was a missed newspaper delivery, so the individual did not see an advertisement for the job counseling program), or 2) even for the same individual type U (and hence with the same potential outcomes), there are sub-types of individuals with differences in whether they choose to participate (e.g. conditional on U, there is heterogeneity in costs of participation).

S9 allows some violations of "deterministic" monotonicity at the individual level (the simultaneous presence of "compliers" and "defiers"), but requires that – conditional on the individual type U – the *probability* of treatment rises when Z moves from 0 to 1. In other words, S9 requires that – conditional on U – on average the "compliers" outnumber the "defiers". To use the notation in the literature, where D_0 and D_1 are the possible treatments when $Z = 0$ or 1, respectively, the monotonicity condition discussed in the literature is $\Pr[D_1 > D_0] = 1$. By contrast, S9 requires $\Pr[D_1 > D_0|U] - \Pr[D_1 < D_0|U] \geq 0$. Integrating over U, S9 thus implies that $\Pr[D_1 > D_0] - \Pr[D_1 < D_0] \geq 0$, but the converse is not true. Furthermore, while $\Pr[D_1 > D_0] = 1$ implies S9, the converse is not true.

It follows that

$$E[y(D, w(u), u)|Z = z, U = u] = y(0, w(u), u) + p_z^*(w(u), u)\Delta(w(u), u)$$

Averaging over the distribution of U conditional on Z yields

$$E[y(D, w(U), U)|Z = z] = \int y(0, w(u), u) + p_z^*(w(u), u)\Delta(w(u), u)\, dF_{U|Z=z}(u)$$

[18] Without the monotonicity condition, the other point of support would be $(1 - p_{z1})$, the latent propensity of the "defiers".

Taking the difference between the $Z=0$ and $Z=1$ individuals, this yields the reduced-form

$$E[Y|Z=1]-E[Y|Z=0]=\int \Delta(w(u),u)\left[p_1^*(w(u),u)-p_0^*(w(u),u)\right]dF_U(u)$$

where D3 allows us to combine the two integrals. Note also that without S8, we would be unable to factor out the term $\Delta(w(u),u)$.

It is useful here to contrast the DGP given by D3 and S8 with the randomized experiment with perfect compliance, in how it confronts the problem posed by Equation (5). With perfect compliance, the randomization made it so that P^* was the same constant for both treated and control individuals, so the two terms in Equation (5) could be combined. With non-random selection into treatment, we must admit the possibility of variability in P^*. But instead of making the contrast between $D=1$ and $D=0$, it is made between $Z=1$ versus $Z=0$ individuals, who, by D3, have the same distribution of types ($F_{U|Z=1}(u)=F_{U|Z=0}(u)$). Thus, the randomized instrument allows us to compare two groups with the same *distribution* of latent propensities P^*: $F_{P^*|Z=1}(p^*)=F_{P^*|Z=0}(p^*)$.

Dividing the preceding equation by a normalizing factor, it follows that the Wald Estimand will identify

$$\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}=\int \Delta(w(u),u)\frac{p_1^*(w(u),u)-p_0^*(w(u),u)}{E[D|Z=1]-E[D|Z=0]}dF_U(u) \tag{9}$$

[19]Therefore, there is an alternative to the interpretation of the Wald estimand as the LATE. It can be viewed as the *weighted* average treatment effect for the entire population where the weights are proportional to the increase in the probability of treatment caused by the instrument, $p_1^*(w(u),u)-p_0^*(w(u),u)$.[20] This weighted average interpretation requires the weights to be non-negative, which will be true if and only if the probabilistic monotonicity condition S9 holds. Note the connection with the conventional LATE interpretation: when treatment is a deterministic function of Z, then the monotonicity means only the compliers (i.e. $p_1^*(w(u),u)-p_0^*(w(u),u)=1$) collectively receive 100 percent of the weight, while all other units receive 0 weight.

The general framework given by Equations (1), (2), (3), and the weaker monotonicity condition S9 thus leads to a less "local" interpretation than LATE. For example, Angrist and Evans (1998) use a binary variable

[19] Note that $\int p_1^*(w(u),u)-p_0^*(w(u),u)\,dF_U(u)=E[D|Z=1]-E[D|Z=0]$.

[20] Alternatively, one can view the weights as being proportional to the fraction of compliers in excess of the defiers among individuals of the same type U: $p_1^*(w(u),u)-p_0^*(w(u),u)=\Pr[D_1>D_0|U=u]-\Pr[D_1<D_0|U=u]$.

that indicates whether the first two children were of the same gender (*Same Sex)* as an instrument for whether the family ultimately has more than 2 children (*More than 2)*. They find a first-stage coefficient of around 0.06. The conventional monotonicity assumption, which presumes that D is a deterministic function of Z, leads to the interpretation that we know that 6 percent of families are "compliers": those that are induced to having a third child because their first two children were of the same gender. This naturally leads to the conclusion that the average effect "only applies" to 6 percent of the population.

In light of Equation (9), however, an alternative interpretation is that the Wald estimand yields a weighted average of 100 percent of the population, with individual weights proportional to the individual-specific impact of (*Same Sex*) on (*More than 2*). In fact, if (*Same Sex*) had the same .06 impact on the probability of having more than 2 children *for all families*, the Wald Estimand will yield the ATE. Nothing in this scenario prevents substantial amount of variation in $p_1^*(w(U),U)$, $p_0^*(w(U),U)$, (and hence P^*), as well as non-random selection into treatment (e.g. correlation between P^* and $y(d,W,U)$).[21] With our hypothetical instrument of "providing information about the job counseling program", a first-stage effect on participation of 0.02 can be interpreted as a 0.02 effect in probability for *all individuals.*

In summary, the data generating process given by D3, S8, and S9 – compared to one where there is deterministic monotonicity – is a broader characterization of the models for which the Wald estimand identifies an average effect. Accordingly, the Wald estimand can have a broader interpretation as a weighted average treatment effect. The framework used to yield the LATE interpretation restricts the heterogeneity in P^* to have only three points of support, 0, p_{z1}, and 1. Thus, the LATE interpretation – which admits that effects can only be identified for those with $P^* = p_{z1}$ is one that most exaggerates the "local" or "unrepresentativesness" of the Wald-identified average effect.

Finally, it is natural to wonder why there is so much of a focus on the Wald estimand. In a purely ex post evaluation analysis, the reason is *not* that IV is a "favorite" or "common" estimand.[22] . Rather, in an ex post evaluation, we may have limited options, based on the realities of how the program was conducted, and what data are available. So, for example, as analysts we may be confronted with an instrument, "provision of information about the job counseling program" (Z), which was indeed randomized as described by D3, and on purely theoretical grounds, we are comfortable with the additional structural assumptions S8 and S9.

[21] But in this example, $p_0^*(w(U),U)$ would have to be bounded above by $1-0.06=0.94$.

[22] Heckman and Vytlacil (2005) andHeckman et al. (2006) correctly observe that from the perspective of the ex ante evaluation problem a singular focus on estimators without an articulated model will not, in general, be helpful in answering a question of economic interest. In an ex post evaluation, however, careful qualification of what parameters are identified from the experiment (as opposed to the parameters of more economic interest) is a desirable feature of the evaluation.

But suppose we are limited by the observable data (Y,D,Z), and know nothing else about the structure given in Equations (1), (2), and (3), and therefore wish our inferences to be invariant to any possible behavioral model consistent with those equations. If we want to identify some kind of average $\Delta(w(U),U)$, then what alternative do we have but the Wald estimand? It is not clear there is one.

The definition of the weights of "interest" is precisely the first step of an *ex ante* evaluation of a program. We argue that the results of an analysis that yields us an average effect, as in (9), may well not be the direct "parameter of interest", but could be used as an ingredient to predict such a parameter in an *ex ante* evaluation analysis. We illustrate this notion with a simple example below.

3.2.1 Assessment

In terms of our three criteria to assess internal validity, how does this research design fare – particularly in comparison to the randomized experiment with perfect compliance? First, only *part* of the data generating process given by D3, S8, and S9 is a literal description of the assignment process: if Z is truly randomly assigned, then D3 is not so much an assumption, but a description. On the other hand, S8 and S9 will typically be conjectures about behavior rather than being an implication of our institutional knowledge.[23] This is an example where there are "gaps" in our understanding of the assignment process, and structural assumptions work together with experimental variation to achieve identification.

As for our second criterion, with the addition of the structure imposed by S8 and S9, it is clear that the class of all behavioral models for which the causal inference in Equation (9) is valid is smaller. It is helpful to consider, for our hypothetical instrument, the kinds of economic models that would or would not be consistent with S8 and S9. If individuals are choosing the best job search activity amongst all known available feasible options, then the instrument of "providing information about the existence of a career counseling program" could be viewed as adding one more known alternative. A standard revealed preference argument would dictate that if an individual already chose to participate under $Z = 0$, then it would still be optimal if $Z = 1$: this would satisfy S9. Furthermore, it is arguably true that most attempts at modeling this process would not specify a direct impact of this added information on human capital; this would be an argument for S8. On the other hand, what if the information received about the program carried a signal of some other factor? It could indicate to the individual, that the state agency is monitoring their job

[23] The exception to this is that, in some cases, our institutional knowledge may lead us to know that those assigned $Z = 0$ are barred from receiving treatment. S9 will necessarily follow.

search behavior more closely. This might induce the individual to search more intensively, independently of participating in the career counseling program; this would be violation of the exclusion restriction S8. Or perhaps the information provided sends a positive signal about the state of the job market and induces the individual to pursue other job search activities instead of the program; this might lead to a violation of S9.

For our third criterion, we can see that some aspects of D3, S8, and S9 are potentially testable. Suppose the elements of W can be categorized into the vector W^- (the variables determined *prior* to Z) and W^+ (after Z). And suppose we can observe a subset of elements from each of these vectors as variables X^- and X^+, respectively. Then the randomization in D3 has the direct implication that the distributions of X^- for $Z = 1$ and $Z = 0$ should be identical:

$$F_{X^-|Z=1}(x) = F_{X^-|Z=0}(x)$$

Furthermore, since the exclusion restriction S8 dictates that all factors W that determine Y are not influenced by Z, then D3 and S8 jointly imply that the distribution of X^+ are identical for the two groups:

$$F_{X^+|Z=1}(x) = F_{X^+|Z=0}(x)$$

The practical limitation here is that this test pre-supposes the researcher's X^+ really do reflect elements of W^+ that influence Y. If X^+ are not a subset of W^+, then even if there is imbalance in X^+, S8 could still hold. Contrast this with the implication of D3 (and D1 and D2) that *any* variable determined prior to the random assignment should have a distribution that is identical between the two randomly assigned groups. Also, there seems no obvious way to test the proposition that Z does not directly impact Y, which is another condition required by S8.

Finally, if S9 holds, it must also be true that

$$\Pr\left[D = 1|X^- = x, Z = 1\right] - \Pr\left[D = 1|X^- = x, Z = 0\right] \geq 0, \ \forall x$$

That is, if probabilistic monotonicity holds for all U, then it must also hold for groups of individuals, defined by the value of X^- (which is a function of U). This inequality also holds for *any* variables determined prior to Z.

In summary, we conclude (unsurprisingly) that for programs where there is random assignment in the "encouragement" of individuals to participate, causal inferences will be of strictly lower internal validity,

relative to the perfect compliance case. Nevertheless, the design does seem to satisfy – even if to a lesser degree – the three criteria that we are considering. Our knowledge of the assignment process does dictate an important aspect of the statistical model (and other aspects need to be modeled with structural assumptions), the causal inferences using the Wald estimand appear valid within a reasonably broad class of models (even it is not as broad as that for the perfect compliance case), and there are certain aspects of the design that generate testable implications.

3.2.2 Ex Ante Evaluation: Extrapolating from LATE to ATE

Perhaps the most common criticism leveled at the LATE parameter is that it may not be the "parameter of interest".[24] By the same token, one may have little reason to be satisfied with the particular weights in the average effect expressed in (9). Returning to our hypothetical example in which the instrument "provide information on career counseling program" is randomized, the researcher may not be interested in an average effect that over-samples those who are more influenced by the instrument. For example, a researcher might be interested in predicting the average impact of individuals of a *mandatory* job counseling program, like the hypothetical example in the case of the randomized experiment with perfect compliance. That is, it may be of interest to predict what would happen if people were required to participate (i.e. every UI claimant will receive a call/visit from a job counselor). Moreover, it has been suggested that LATE is an "instrument-specific" parameter and a singular focus on LATE risks conflating "definition of parameters with issues of identification" (Heckman and Vytlacil, 2005): different instruments can be expected to yield different "LATEs".

Our view is that these are valid criticisms from the perspective of an ex ante evaluation standpoint, where causal "parameters of interest" are defined by a theoretical framework describing the policy problem. But from an ex post evaluation perspective, within which internal validity is the primary goal, these issues are by definition unimportant. When the goal is to describe whatever one can about the causal effects of a program that was actually implemented (e.g. randomization of "information about the job counseling program", Z), a rigorous analysis will lead to precise statements about the causal phenomena that are *possible* to credibly identify. Sometimes what one *can* credibly identify may correspond to a desired parameter from a well-defined ex ante evaluation; sometimes it will not.

[24] Discussions on this point can be found, for example, in Heckman and Vytlacil (2001b); Heckman (2001); Heckman and Vytlacil (2005)), as well as Heckman and Vytlacil (2007a,b); Abbring and Heckman (2007).

Although there has been considerable emphasis in the applied literature on the fact that LATE may differ from ATE, as well as discussion in the theoretical literature about the "merits" of LATE as a parameter, far less effort has been spent in actually using estimates of LATE to learn about ATE. Even though standard "textbook" selection models can lead to simple ways to extrapolate from LATE to ATE (see Heckman and Vytlacil (2001a); Heckman et al. (2001, 2003)), our survey of the applied literature revealed very few other attempts to actually produce these extrapolations.

Although we have argued that the ATE from a randomized experiment may not directly correspond to the parameter of interest, for the following derivations, let us stipulate that ATE is useful, either for extrapolation (as illustrated in Section 3.1.4), or as an "instrument-invariant" benchmark that can be compared to other ATEs extrapolated from other instruments or alternative identification strategies.

Consider re-writing the structure given by Equations (1), (2), and (3), and D3, S8, and S9 as

$$\begin{aligned} Y_1 &= \mu_1 + U_1 \\ Y_0 &= \mu_0 + U_0 \\ D &= 1\left[Zg_1 + (1-Z)\,g_0 + U^D \geq 0\right] \\ Y &= DY_1 + (1-D)\,Y_0 \end{aligned}$$

where μ_1, μ_0 are constants, Z is a binary instrument, and (U_1, U_0) characterize both the individual's type and all other factors that determine Y, and is independent of Z by D3 and S8. g_1 and g_0 are constants in the selection equation: S9 is satisfied. U_1, U_0, U^D can be normalized to be mean zero error terms. The ATE is by construction equal to $\mu_1 - \mu_0$.

Let us adopt the following functional form assumption:

- S10 (Normality of Errors): $\left(U_1, U^D\right)$ and $\left(U_0, U^D\right)$ are both bivariate normals with covariance matrices $\begin{pmatrix} \sigma_1 & \rho_1\sigma_1 \\ \rho_1\sigma_1 & 1 \end{pmatrix}$ and $\begin{pmatrix} \sigma_0 & \rho_0\sigma_0 \\ \rho_0\sigma_0 & 1 \end{pmatrix}$, respectively, where we are – without loss of generality – normalizing U^D to have a variance of 1.

This is simply the standard dummy endogenous variable system (as in Heckman (1976, 1978); Maddala (1983)), with the special case of a dummy variable instrument, and is a case that is considered in recent work by (Angrist (2004) and Oreopoulos (2006)). With this one functional form assumption, we obtain a

relationship between LATE and ATE.

In particular we know that

$$LATE = \mu_1 - \mu_0 + (\rho_1\sigma_1 - \rho_0\sigma_0)\frac{\phi(-g_1) - \phi(-g_0)}{\Phi(-g_0) - \Phi(-g_1)} \tag{10}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal, respectively.[25] This is a standard result that directly follows from the early work on selection models (Heckman (1976, 1978) See also Heckman et al. (2001, 2003)). This framework has been used to discuss the relationship between ATE and LATE (seeHeckman et al. (2001), Angrist (2004), and Oreopoulos (2006)). With a few exceptions (such as Heckman et al. (2001), for example), other applied researchers typically do not make use the fact that even with information on just the three variables Y, D, and Z, the "selection correction" term in Equation (10) can be computed. In particular

$$\begin{aligned} E[Y|D=1,Z=z] &= \mu_1 + \rho_1\sigma_1\left(\frac{\phi(-g_z)}{1-\Phi(-g_z)}\right) \\ E[Y|D=0,Z=z] &= \mu_0 - \rho_0\sigma_0\left(\frac{\phi(-g_z)}{\Phi(-g_z)}\right) \end{aligned}$$

implies that

$$\rho_1\sigma_1 = \frac{E[Y|D=1,Z=1] - E[Y|D=1,Z=0]}{\frac{\phi(\Phi^{-1}(1-E[D|Z=1]))}{E[D|Z=1]} - \frac{\phi(\Phi^{-1}(1-E[D|Z=0]))}{E[D|Z=0]}} \tag{11}$$

and analogously that

$$\rho_0\sigma_0 = \frac{E[Y|D=0,Z=0] - E[Y|D=0,Z=1]}{\frac{\phi(\Phi^{-1}(1-E[D|Z=1]))}{1-E[D|Z=1]} - \frac{\phi(\Phi^{-1}(1-E[D|Z=0]))}{1-E[D|Z=0]}} \tag{12}$$

. Having identified $\rho_1\sigma_1$ and $\rho_0\sigma_0$, we have the expression

$$ATE = LATE - (\rho_1\sigma_1 - \rho_0\sigma_0)\cdot\frac{\phi\left(\Phi^{-1}(1-E[D|Z=1])\right) - \phi\left(\Phi^{-1}(1-E[D|Z=0])\right)}{E[D|Z=1] - E[D|Z=0]} \tag{13}$$

.

[25] To see this, note that $E[Y|Z=1] = E[Y_1|D=1,Z=1]\Pr[D=1|Z=1] + E[Y_0|D=0,Z=1]\Pr[D=0|Z=1]$ which is equal to $\left(\mu_1 + \rho_1\sigma_1\frac{\phi(-g_1)}{1-\Phi(-g_1)}\right)(1-\Phi(-g_1)) + \left(\mu_0 - \rho_0\sigma_0\frac{\phi(-g_1)}{\Phi(-g_1)}\right)(\Phi(-g_1))$. We can decompose the first term into two terms to yield $\left(\mu_1 + \rho_1\sigma_1\frac{\phi(-g_0)}{1-\Phi(-g_0)}\right)(1-\Phi(-g_0)) + (\mu_1(\Phi(-g_0) - \Phi(-g_1)) + \rho_1\sigma_1(\phi(-g_1) - \phi(-g_0))) + \left(\mu_0 - \rho_0\sigma_0\frac{\phi(-g_1)}{\Phi(-g_1)}\right)(\Phi(-g_1))$. Taking the difference between this and an analogous expression for $E[Y|Z=0]$, the first and fourth terms cancel. Dividing the result by $\Phi(-g_0) - \Phi(-g_1)$ yields Equation (10).

This expression is quite similar to that given in Section 3.1.4. Once again, the result of an ex post evaluation can be viewed as the leading term in the extrapolative goal of an ex ante evaluation: to obtain the effects of a program that was *not* implemented (i.e. random assignment with perfect compliance) or the ATE. This expression also shows how the goals of the ex post and ex ante evaluation problems can be complementary. Ex post evaluations aim to get the best estimate of the first term in the above equation, whereas ex ante evaluations are concerned with the assumptions necessary to extrapolate from LATE to ATE as in the above expression.

LATE versus ATE in Practice

If S10 were adopted, how much might estimates of LATE differ from those of ATE in practice? To investigate this, we obtained data from a select group of empirical studies and computed both the estimates of LATE and the "selection correction" term in (13), using (11) and (12).[26]

The results are summarized in Table 1. For each of the studies, we present the simple difference in the means $E[Y|D=1]-E[Y|D=0]$, the Wald estimate of LATE, the implied selection error correlations, $\rho_1\sigma_1$, $\rho_0\sigma_0$, a selection correction term, and the implied ATE. For comparison, we also give the average value of Y. Standard deviations are in brackets. The last row of the table gives the value of the second term in (13), and its significance (calculated using the delta method).

The quantity $\rho_1\sigma_1-\rho_0\sigma_0$ is the implied covariance between the gains U_1-U_0 and the selection error U^D. A "selection on gains" phenomenon would imply $\rho_1\sigma_1-\rho_0\sigma_0>0$. With the study of Abadie et al. (2002)in the first column, we see a substantial negative selection term, where the resulting LATE is significantly less than either the simple difference, or the implied ATE. If indeed the normality assumption is correct, this would imply that for the purposes of obtaining ATE, which might be the target of an ex ante evaluation, simply using LATE would be misleading, and ultimately even worse than using the simple difference in means.

On the other hand, in the analysis of Angrist and Evans (1998), the estimated LATE is actually quite similar to the implied ATE. So while a skeptic might consider it "uninteresting" to know the impact of having more than 2 children for the "compliers" (those whose family size was impacted by the gender mix

[26] We chose the studies for this exercise in the following way. We searched articles mentioning "local average treatment effect(s)", as well as articles which cite Imbens and Angrist (1994) and Angrist et al. (1996). We restricted the search to articles that are published in *American Economic Review*, *Econometrica*, *Journal of Political Economy*, or *Quarterly Journal of Economics*, or *Review of Economic Studies*. From this group, we restricted our attention to studies in which both the instrument and the treatment were binary. The studies presented in the table are the ones in this group for which we were able to obtain the data, successfully replicate the results, and where computing the IV estimate without covariates did not substantially influence the results (Angrist and Evans, 1998; Abadie et al., 2002; Angrist et al., 2006; Field, 2007).

of the first two children), it turns out in this context – if one accepts the functional form assumption – LATE and ATE do not differ very much.[27] The other studies are examples of intermediate cases: LATE may not be equal to ATE, but it is closer to ATE than the simple difference $E[Y|D=1]-E[Y|D=0]$.

Our point here is neither to recommend nor to discourage the use of this normal selection model for extrapolation. Rather, it is to illustrate and emphasize that even if LATE is identifying an average effect for a *conceptually* different population from the ATE, this does not necessarily mean that the two quantities in an actual application are very different. Our other point is that any inference that uses LATE to make *any* statement about the causal phenomena outside the context from which a LATE is generated, must necessarily rely on a structural assumption, whether implicitly or explicitly. In this discussion of extrapolating from LATE to ATE, we are being explicit that we are able to do this through a bivariate normal assumption. While such an assumption may seem unpalatable to some, it is clear that to insist on making no extrapolative assumptions is to abandon the ex ante evaluation goal entirely.

3.3 Regression Discontinuity Design: Sharp

This section provides an extended discussion of identification and estimation of the regression discontinuity (RD) design. RD designs were first introduced by Thistlethwaite and Campbell (1960) as a way of estimating treatment effects in a non-experimental setting, where treatment is determined by whether an observed "assignment" variable (also referred to in the literature as the "forcing" variable or the "running" variable) exceeds a known cutoff point. In their initial application of RD designs, Thistlethwaite and Campbell (1960) analyzed the impact of merit awards on future academic outcomes, using the fact that the allocation of these awards was based on an observed test score. The main idea behind the research design was that individuals with scores just below the cutoff (who did not receive the award) were good comparisons to those just above the cutoff (who did receive the award). Although this evaluation strategy has been around for almost fifty years, it did not attract much attention in economics until relatively recently.

Since the late 1990s, a growing number of studies have relied on RD designs to estimate program effects in a wide variety of economic contexts. Like Thistlethwaite and Campbell (1960), early studies by Van der Klaauw (2002) and Angrist and Lavy (1999) exploited threshold rules often used by educational institutions

[27] The obvious problem with the functional form here is that "Worked for pay" is a binary variable. One can still use the bivariate normal framework as an approximation, if Y is interpreted to be the latent probability of working, which can be continuously distributed (but one has to ignore the fact that the tails of the normal necessarily extend beyond the unit interval). In this case, the "average" effect is the average effect on the underlying probability of working.

to estimate the effect of financial aid and class size, respectively, on educational outcomes. Black (1999) exploited the presence of discontinuities at the geographical level (school district boundaries) to estimate the willingness to pay for good schools. Following these early papers in the area of education, the past five years have seen a rapidly growing literature using RD designs to examine a range of questions. Examples include: the labor supply effect of welfare, unemployment insurance, and disability programs; the effects of Medicaid on health outcomes; the effect of remedial education programs on educational achievement; the empirical relevance of median voter models; and the effects of unionization on wages and employment.

An important impetus behind this recent flurry of research is a recognition, formalized by Hahn et al. (2001), that RD designs require seemingly mild assumptions compared to those needed for other non-experimental approaches. Another reason for the recent wave of research is the belief that the RD design is not "just another" evaluation strategy, and that causal inferences from RD designs are potentially more credible than those from typical "natural experiment" strategies (e.g. difference-in-differences or instrumental variables), which have been heavily employed in applied research in recent decades. This notion has a theoretical justification: Lee (2008) formally shows that one need not *assume* the RD design isolates treatment variation that is "as good as randomized"; instead, such randomized variation is a *consequence* of agents' inability to precisely control the assignment variable near the known cutoff.

So while the RD approach was initially thought to be "just another" program evaluation method with relatively little general applicability outside of a few specific problems, recent work in economics has shown quite the opposite.[28] In addition to providing a highly credible and transparent way of estimating program effects, RD designs can be used in a wide variety of contexts covering a large number of important economic questions. These two facts likely explain why the RD approach is rapidly becoming a major element in the toolkit of empirical economists.

Before presenting a more formal discussion of various identification and estimation issues, we first briefly highlight what we believe to be the most important points that have emerged from the recent theoretical and empirical literature on the RD design.[29] In this chapter, we will use V to denote the assignment variable, and treatment will be assigned to individuals when V exceeds a known threshold c, which we later normalize to 0 in our discussion.

[28] See Cook (2008) for an interesting history of the RD design in education research, psychology, statistics, and economics. Cook argues the resurgence of the RD design in economics is unique as it is still rarely used in other disciplines.

[29] Recent surveys of the RD design in theory and practice include Lee and Lemieux (2009), Van der Klaauw (2008a), and Imbens and Lemieux (2008a).

- **RD designs can be invalid if individuals can precisely manipulate the "assignment variable".**

  When there is a payoff or benefit to receiving a treatment, it is natural for an economist to consider how an individual may behave to obtain such benefits. For example, if students could effectively "choose" their test score V through effort, those who chose a score c (and hence received the merit award) could be somewhat different from those who chose scores just below c. The important lesson here is that the existence of a treatment being a discontinuous function of an assignment variable is *not* sufficient to justify the validity of an RD design. Indeed, if anything, discontinuous rules may generate incentives, causing behavior that would *invalidate* the RD approach.

- **If individuals – even while having some influence – are unable to *precisely* manipulate the assignment variable, *a consequence* of this is that the variation in treatment near the threshold is randomized as though from a randomized experiment.**

  This is a crucial feature of the RD design, and a reason that RD designs are often so compelling. Intuitively, when individuals have imprecise control over the assignment variable, even if some are especially likely to have values of V *near* the cutoff, *every* individual will have approximately the same probability of having an V that is just above (receiving the treatment) or just below (being denied the treatment) the cutoff – similar to a coin-flip experiment. This result clearly differentiates the RD and IV (with a non-randomized instrument) approaches. When using IV for causal inference, one must *assume* the instrument is exogenously generated as if by a coin-flip. Such an assumption is often difficult to justify (except when an actual lottery was run, as inHearst et al. (1986)or Angrist (1990), or if there were some biological process, e.g. gender determination of a baby, mimicking a coin-flip). By contrast, the variation that RD designs isolate is randomized *as a consequence* of the assumption that individuals have imprecise control over the assignment variable (Lee, 2008).

- **RD designs can be analyzed – and tested – like randomized experiments.**

  This is the key implication of the local randomization result. If variation in the treatment near the threshold is approximately randomized, then it follows that all "baseline characteristics" – all those variables determined prior to the realization of the assignment variable – should have the same distribution just above and just below the cutoff. If there is a discontinuity in these baseline covariates, then at a minimum, the underlying identifying assumption of individuals' inability to precisely manipulate the assignment variable is unwarranted. Thus, the baseline covariates are used to *test* the validity of

the RD design. By contrast, when employing an IV or a matching/regression-control strategy in non-experimental situations, assumptions typically need to be made about the relationship of these other covariates to the treatment and outcome variables.[30]

- **The treatment effects from RD can be interpreted as a weighted average treatment effect.**
  It is tempting to conclude that the RD delivers treatment effects that "only apply" for the sub-population of individuals whose V is arbitrarily close to the threshold c. Such an interpretation would imply that the RD identifies treatment effects for "virtually no one". Fortunately, as we shall see below, there is an alternative interpretation: the average effect identified by a valid RD is that of a weighted average treatment effect where the weights are the relative ex ante probability that the value of an individual's assignment variable will be in the neighborhood of the threshold (Lee, 2008).

Randomized Experiments from Non-random Selection

As argued in Lee and Lemieux (2009), while there are some mechanical similarities between the RD design and a "matching on observables" approach or between the RD design and an instrumental variables approach, the RD design can instead be viewed as a close "cousin" of the randomized experiment, in the sense that what motivates the design and what "dictates" the modeling is specific institutional knowledge of the treatment assignment process. We illustrate this by once again using the common framework given by Equations (1), (2), and (3). We begin with the case of the "sharp" RD design, whereby the treatment status is a deterministic "step-function" of an observed assignment variable V. That is, $D = 1$ if and only if V crosses the discontinuity threshold (normalized to 0 in our discussion).

Returning to our hypothetical job search assistance program, suppose that the state agency needed to ration the number of participants in the program, and therefore mandated treatment (personal visit and/or phone call from job counselor) for those whom the agency believed would the program would most greatly benefit. In particular, suppose the agency used information on individuals' past earnings and employment information generate a score V that indicated the likely benefit of the program to the individual, and determined treatment status based on whether that V exceeded the threshold 0.

Such a discontinuous rule can be described as

- D4: (Discontinuous rule) $D = 1[V \geq 0]$: This implies that Equation (2) becomes $P^* = p^*(W,U) = \Pr[V \geq 0|W,U]$.

[30] Typically, one assumes that *conditional on the covariates*, the treatment (or instrument) is "as good as" randomly assigned.

- D5: V is observed.

Both D4 and D5 come from institutional knowledge of how treatment is assigned, and thus are more descriptions ("D") than assumptions.

We further assume that

- S11: (Positive density at the threshold): $f_V(0) > 0$.

This assumption ensures that there are some individuals at the threshold.

Since we have introduced a variable V that is realized before D, we must specify its relation to Y, so that Equation (3) becomes

- S12 (Continuous impact of V) $Y = y(D,W,U,V)$, where $y(d,w,u,v)$ is continuous in v (at least in a neighborhood of $v = 0$).

This assumption states that for any individual type U, as V crosses the discontinuity threshold 0, any change in Y must be attributable to D and D only. As we shall see, this assumption, while necessary, is *not* sufficient for the RD to deliver valid causal inferences.

The most important assumption for identification is

- S13 (Continuous Density; Incomplete/Imprecise Control of V) The distribution of V conditional on U has density $f_{V|U}(v)$ that is continuous in v, at least in the neighborhood of $v = 0$.

This condition, which we will discuss in greater detail below, says that individuals – no matter how much they can influence the *distribution* of V with their actions – cannot *precisely* control V, even if they may make decisions W in anticipation of this uncertainty.

There are at least two alternative interpretations of this condition. One is that individuals may actually precisely control V, but they are responding to different external factors, which generates a distribution of different possible Vs that could occur depending on these outside forces. S13 says that the distribution of V – as driven by those outside forces – must have a continuous density. The other interpretation is that for each unobservable type U, there are "sub-types", where each sub-type chooses a different level of V. In this case, S13 is a statement about the distribution of V (as generated by the heterogeneity in "sub-types") being continuous conditional on the type U.

As Lee (2008) shows, it is precisely S13 that will generate a local randomization result. In particular, S13 implies

$$\begin{aligned}\lim_{v\uparrow 0} f_{U|V=v}(u) &= \lim_{v\uparrow 0} \frac{f_{V|U=u}(v)}{f_V(v)} \cdot f_U(u) \\ &= \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u) \\ &= f_{U|V=0}(u)\end{aligned}$$

which says that the distribution of the unobserved "types" U will be approximately equal on either side of the discontinuity threshold in a neighborhood of 0. This is the sense in which it accomplishes *local randomization*, akin to the randomization in an experiment. The difference is in how the problem expressed in Equation (5) is being confronted. The experimenter in the randomized experiment ensures that treated and non-treated individuals have the same distribution of latent propensities P^* by dictating that all individuals have the same fixed $P^* = p_0$. In the non-experimental context here, we have no control over the distribution of P^*, but if S13 holds, then the (non-degenerate) distribution of P^* (which is a function of U) will approximately be equal between treated and non-treated individuals – for those with realized V in a small neighborhood of 0.

Now consider the expectation of Y at the discontinuity threshold

$$\begin{aligned}E[Y|V=0] &= E[y(1,w(U),U,0)|V=0] \\ &= \int y(1,w(u),u,0) f_{U|V=0}(u)\,du \\ &= \int y(1,w(u),u,0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u)\,du\end{aligned}$$

where the third line follows from Bayes' Rule. Similarly, we have

$$\begin{aligned}\lim_{v\uparrow 0} E[Y|V=v] &= \lim_{v\uparrow 0} E[y(0,w(U),U,v)|V=v] \\ &= \lim_{v\uparrow 0} \int y(0,w(u),u,v) \frac{f_{V|U=u}(v)}{f_V(v)} \cdot f_U(u)\,du \\ &= \int y(0,w(u),u,0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u)\,du\end{aligned}$$

where the last line follows from the continuity assumption S12 and S13.

The RD estimand – the difference between the above two quantities – is thus

$$\begin{aligned} E[Y|V=0] - \lim_{v \uparrow 0} E[Y|V=v] &= \int [y(1,w(u),u,0) - y(0,w(u),u,0)] \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u)\, du \quad (14) \\ &= \int \Delta(w(u),u,0) \frac{f_{V|U=u}(0)}{f_V(0)} \cdot f_U(u)\, du \end{aligned}$$

. That is, the discontinuity in the conditional expectation function $E[Y|V=v]$ identifies a *weighted* average of causal impacts $\Delta(w(u),u,0)$, where the weights are proportional to type U's relative likelihood that the assignment variable V is in the neighborhood of the discontinuity threshold 0.

Equation (14) provides a quite different alternative to the interpretation of the estimand as "the treatment effect for those whose V are close to zero" – which connotes a very limited inference, because in the limit, there are no individuals at $V=0$. The variability in weights in (14) depend very much on the typical scale of $f_{V|U}(\cdot)$ relative to the location of $f_{V|U}(\cdot)$ across the U types. That is, if for each type U there is negligible variability in V, then the RD estimand will indeed identify a treatment effect for those individuals who can be most expected to have V close to the threshold. On the other extreme, if there is large variability, with $f_{V|U}(\cdot)$ having flat tails, the weights will tend to be more uniform.[31]

One of the reasons why RD design can be viewed as a "cousin" of the randomized experiment is that the latter is really a special case of the sharp RD design. When randomly assigning treatment, one can imagine accomplishing this through a continuously distributed random variable V that has the same distribution for every individual. In that case, V would not enter the function determining Y, and hence S12 would be unnecessary. It would follow that $f_{V|U}(v) = f_V(v)$, and consequently every individual would receive equal weight in the average effect expression in Equation (14).

The final point to notice about the average effect that is identified by the RD design is that it is a weighted average of $\Delta(w(U),U,0)$, which may be different from a weighted average of $\Delta(w(U),U,v)$ for some other v, even if the weights $\frac{f_{V|U=u}(0)}{f_V(0)}$ do not change. Of course, there is no difference between the two quantities in situations where V is thought to have no impact on the outcome (or the individual treatment effect). For example, if a test score V is used for one and only one purpose – to award a scholarship D – then it might be reasonable to assume that V has no other impact on future educational outcomes Y. As discussed in Lee (2008) and Lee and Lemieux (2009), in other situations, the concept of $\Delta(w(U),U,v)$ for

[31] For example, for the uniform density $f_{V|U}(v) = \frac{1}{\theta} \cdot 1[\mu_U \leq v \leq \mu_U + \theta]$, the weights would be identical across types, even though there would be variability in the probability of treatment driven by variability in μ_U.

values of v other than 0 may not make much practical sense. For example, in the context of estimating the electoral advantage to incumbency in U.S. House elections (Lee, 2008), it is difficult to conceive of the counterfactual $y(0, w(u), u, v)$ when v is away from the threshold: what does it mean for the outcome that *would* have been obtained if the candidate who became the incumbent with 90 percent of the vote had *not* become the incumbent, having won 90 percent of the vote? Here, incumbent status is defined by V.

3.3.1 Assessment: Valid or Invalid RD?

We now assess this design on the basis of the three criteria discussed in Section2.2.1. First, some of the conditions for identification are indeed literally descriptions of the assignment process: D4 and D5. Others, like S11 through S13, are conjectures about the assignment process, and the underlying determinants of Y. S11 requires that there is positive density of V at the threshold. S12 allows V – which can be viewed as capturing "all other factors" that determine D - - to have its own structural effect on Y. It is therefore not as restrictive as a standard exclusion restriction, but S12 does require that the impact of V is continuous. S13 is the most important condition for identification, and we discuss it further below.

Our second criterion is the extent to which inferences could be consistent with many competing behavioral models. The question is to what extent does S12 and S13 restrict the class of models for which the RD causal inference remains valid? When program status is determined solely on the basis of a score V, and V is used for nothing else but the determination of D, we expect most economic models to predict that the only reason why V would have a discontinuous impact on Y would be because an individual's status switches from non-treated to treated. So, from the perspective of modeling economic behavior, S12 does not seem to be particularly restrictive.

By contrast, S13 is potentially restrictive, ruling out some plausible economic behavior. Are individuals able to influence the assignment variable, and if so, what is the nature of this control? This is probably the most important question to ask when assessing whether a particular application should be analyzed as an RD design. If individuals have a great deal of control over the assignment variable and if there is a perceived benefit to a treatment, one would certainly expect individuals on one side of the threshold to be systematically different from those on the other side.

Consider the test-taking example from Thistlethwaite and Campbell (1960). Suppose there are two types of students: A and B. Suppose type A students are more able than B types, and that A types are also keenly aware that passing the relevant threshold (50 percent) will give them a scholarship benefit, while B types

are completely ignorant of the scholarship and the rule. Now suppose that 50 percent of the questions are trivial to answer correctly, but due to random chance, students will sometimes make careless errors when they initially answer the test questions, but would certainly correct the errors if they checked their work. In this scenario, only type *A* students will make sure to check their answers before turning in the exam, thereby assuring themselves of a passing score. The density of their score is depicted in the truncated density in Figure 1. Thus, while we would expect those who barely passed the exam to be a mixture of type *A* and type *B* students, those who barely failed would exclusively be type *B* students. In this example, it is clear that the marginal failing students do *not* represent a valid counterfactual for the marginal passing students. Analyzing this scenario within an RD framework would be inappropriate.

On the other hand, consider the same scenario, except assume that questions on the exam are *not* trivial; there are no guaranteed passes, no matter how many times the students check their answers before turning in the exam. In this case, it seems more plausible that among those scoring near the threshold, it is a matter of "luck" as to which side of the threshold they land. Type *A* students can exert more effort – because they know a scholarship is at stake – but they do not know the exact score they will obtain. This can be depicted by the untruncated density in Figure 1 . In this scenario, it would be reasonable to argue that those who marginally failed and passed would be otherwise comparable, and that an RD analysis *would* be appropriate and would yield credible estimates of the impact of the scholarship.

These two examples make it clear that one must have some knowledge about the mechanism generating the assignment variable, beyond knowing that if it crosses the threshold, the treatment is "turned on". The"folk wisdom" in the literature is to judge whether the RD is appropriate based on whether individuals could *manipulate* the assignment variable and *precisely* "sort" around the discontinuity threshold. The key word here should be "precise", rather than "manipulate". After all, in both examples above, individuals do exert some control over the test score. And indeed in virtually every known application of the RD design, it is easy to tell a plausible story that the assignment variable is to some degree influenced by *someone*. But individuals will not always have *precise* control over the assignment variable. It should, perhaps, seem obvious that it is necessary to rule out *precise* sorting to justify the use of an RD design. After all, individual self-selection into treatment or control regimes is exactly why simple comparison of means is unlikely to yield valid causal inferences. Precise sorting around the threshold is self-selection.

Finally, the data generating process given by D4, D5, S11, S12, and S13 has many testable implications. D4 and D5 are directly verifiable, and S11 can be checked since the marginal density $f_V(0)$ can be observed

from the data. S12 appears fundamentally unverifiable, but S13, which generates the local randomization result, is testable in two ways. First, as McCrary (2008) points out, the continuity of each type's density $f_{V|U}(\cdot)$ (S13) implies that the observed marginal density $f_V(\cdot)$ is continuous, leading to a natural test: examining the data for a potential discontinuity in the density of V at the threshold; McCrary (2008) proposes an estimator for this test. Second, the local randomization result implies that

$$\lim_{v \uparrow 0} F_{W|V=v}(w) = F_{W|V=0}(w)$$

because W is by definition determined prior to D (Equation (1)). This is analogous to the test of randomization where treated and control observations are tested for balance in observable elements of W.[32]

In summary, based on the above three criteria, the RD design does appear to have potential to deliver highly credible causal inferences because some important aspects of the model are literal descriptions of the assignment process and because the conditions for identification are consistent with a broad class of behavioral models (e.g. V can be endogenous – as it is influenced by actions in W – as long there is "imprecise" manipulation of V). Perhaps most importantly, the key condition that is *not* derived from our institutional knowledge (and is the key restriction on economic behavior (S13)), has a testable implication that is as strong as that given by the randomized experiment.

3.3.2 Ex Ante Evaluation: Extrapolating from the RD to Treatment on the Treated

Consider again our hypothetical job search assistance program, where individuals were assigned to the treatment group if their score V (based on earnings and employment information and the state agency's model of who would most benefit from the program) exceeded the threshold 0. From an ex ante evaluation perspective, one could potentially be interested in predicting the policy impact of "shutting down the program" Heckman and Vytlacil (2005). As pointed out in Heckman and Vytlacil (2005), the treatment on the treated parameter (TOT) is an important ingredient for such a prediction. But as we made precise in the ex post evaluation discussion, the RD estimand identifies a particular weighted average treatment effect that in general will be different from the TOT. Is there a way to extrapolate from the average effect in (14) to TOT?

We now sketch out one such proposal for doing this, recognizing this is a relatively uncharted area of research (and that TOT, while an ingredient in computing the impact of the policy of "shutting down the

[32] Additionally, *any* variable determined prior to D – whether or not they are an element of W – should have the same distribution on either side of the discontinuity threshold. This, too, is analogous to the case of randomized assignment.

program", may not be of interest for a different proposed policy). Using D4, D5, S11, S12, and S13, we can see that the TOT is

$$\begin{aligned} E\left[\Delta\left(w\left(U\right),U,V\right)|D=1\right] &= E\left[y\left(1,w\left(U\right),U,V\right)-y\left(0,w\left(U\right),U,V\right)|D=1\right] \\ &= E\left[Y|D=1\right]-E\left[y\left(0,w\left(U\right),U,V\right)|V>0\right] \end{aligned}$$

This reveals that the key missing ingredient is the second term of this difference.

Note that we have

$$\begin{aligned} E\left[y\left(0,w\left(U\right),U,V\right)|V\geq 0\right] &= \frac{1}{1-F_V\left(0\right)}\int_{v\geq 0}\left[\int y\left(0,w\left(u\right),u,v\right)dF_{U|V=v}\left(u\right)\right]f_V\left(v\right)dv \\ &= \frac{1}{1-F_V\left(0\right)}\int_{v\geq 0}\left[\int y\left(0,w\left(u\right),u,v\right)\frac{f\left(v|u\right)}{f\left(v\right)}dF_U\left(u\right)\right]f_V\left(v\right)dv \end{aligned}$$

where the second line again follows from Bayes' rule.

Suppose V has bounded support, and also assume

- S14: (Differentiable Density) Let $f_{V|U=u}\left(v\right)$ have continuous qth derivative on $v\geq 0$, for all u.
- S15: (Differentiable Outcome Function) Let $y\left(0,w\left(u\right),u,v\right)$ have continuous qth derivative on $v\geq 0$, for all u.

With the addition of S14 and S15, this will imply that we can use the Taylor approximation

$$\begin{aligned} E\left[y\left(0,w\left(U\right),U,v\right)|V=v\right] &\approx E\left[y\left(0,w\left(U\right),U,0\right)|V=0\right]+\left.\frac{\partial E\left[y\left(0,w\left(U\right),U,v^*\right)|V=v^*\right]}{\partial v^*}\right|_{v^*=0}v+\ldots \\ &\quad +\frac{1}{q!}\left.\frac{\partial^q E\left[y\left(0,w\left(U\right),U,v^*\right)|V=v^*\right]}{\partial v^{*q}}\right|_{v^*=0}v^q \\ &\approx \lim_{v^*\uparrow 0}E\left[Y|V=v^*\right]+\lim_{v^*\uparrow 0}\left.\frac{\partial E\left[Y|V=v^*\right]}{\partial v^*}\right|_{v^*=0}v+\ldots \\ &\quad +\frac{1}{q!}\lim_{v^*\uparrow 0}\left.\frac{\partial^q E\left[Y|V=v^*\right]}{\partial v^{*q}}\right|_{v^*=0}v^q. \end{aligned}$$

In principle, once this function can be (approximately) identified, one can average the effects over the treated population using the conditional density $f_{V|V\geq 0}\left(v\right)$.

Once again, we see that the leading term of an extrapolation for an ex ante evaluation, is related to the results of an ex post evaluation: it is precisely the counterfactual average in Equation (14). There are

a number of ways of estimating derivatives nonparametrically (Fan and Gijbels, 1996; Pagan and Ullah, 1999), or one could alternatively attempt to approximate the function $E[Y|V=v]$ for $D<0$ with a low-order global polynomial. We recognize that, in practical empirical applications, estimates of higher order derivatives may be quite imprecise.

Nevertheless, we provide this simple method to illustrate the separate role of conditions for credible causal inference (D4, D5, S11, S12, and S13), and the additional structure needed (S14 and S15) to identify a policy-relevant parameter such as the TOT. What alternative additional structure could be imposed on Equations (1), (2), and (3) to identify parameters such as the TOT seems to be an open question, and likely to be somewhat context-dependent.

3.3.3 Estimation Issues for the RD Design

There has been a recent explosion of applied research utilizing the RD design (see Lee and Lemieux (2009)). From this applied literature, a certain "folk wisdom" has emerged about sensible approaches to implementing the RD design in practice. The key challenge with the RD design is how to use a finite sample of data on Y and V to estimate the conditional expectations in the discontinuity $E[Y|V=0]-\lim_{v\uparrow 0}E[Y|V=v]$. Lee and Lemieux (2009) discuss these common practices and their justification in greater detail. Here we simply highlight some of the key points of that review, and then conclude with the recommended "checklist" suggested by Lee and Lemieux (2009).[33]

- **Graphical presentation of an RD design is helpful and informative, but the visual presentation should not be tilted toward either finding an effect or finding no effect.**

  It has become standard to summarize RD analyses with a simple graph showing the relationship between the outcome and assignment variable. This has several advantages. The presentation of the "raw data" enhances the transparency of the research design. A graph can also give the reader a sense of whether the "jump" in the outcome variable at the cutoff is unusually large compared to the bumps in the regression curve away from the cutoff. Also, a graphical analysis can help identify why different functional forms give different answers, and can help identify outliers, which can be a problem in any empirical analysis. The problem with graphical presentations, however, is that there is some room for the researcher to construct graphs making it seem as though there are effects when

[33] See Imbens and Lemieux (2008b)and Van der Klaauw (2008b) for other surveys.

there are none, or hiding effects that truly exist. A way to guard against this visual bias is to partition V into intervals – with the discontinuity threshold at one of the boundaries – and present the mean within each interval. Often, the data on V will already be discrete. This way, the behavior of the function around the threshold is given no special "privilege" in the presentation, yet it allows for the data to "speak for itself" as to whether there is an important jump at the threshold.[34]

- **Non-parametric estimation does not represent a unique, always-preferred"solution" to functional form issues raised by RD designs. It is therefore helpful to view it as a complement to – rather than a substitute for – parametric estimation.**

  Here it is helpful to keep distinct the notions of identification and estimation. The RD design, as discussed above, is non-parametrically identified, and no parametric restrictions are needed to compute $E[Y|V=0] - \lim_{v \uparrow 0} E[Y|V=v]$, given an infinite amount of data. But with a finite sample, one has a choice of different statistics, some referred to as "non-parametric" (e.g. kernel regression, local linear regression), while others considered "parametric" (e.g. a low-order polynomial). As Powell (1994) points out, it is perhaps more helpful to view *models* rather than particular statistics as "parametric" or "non-parametric".[35] The bottom line is that when the analyst chooses a particular functional form (say, a low-order polynomial) in estimation, and the true function does not belong to that polynomial class, then the resulting estimator will, in general, be biased. When the analyst uses a non-parametric procedure such as local linear regression – essentially running a regression using only data points "close" to the cutoff – there will also be bias.[36] With a finite sample, it is impossible to know which case has a smaller bias without knowing something about the true function. There will be some functions where a low-order polynomial is a very good approximation and produces little or no bias, and therefore it is efficient to use all data points – both "close to" and "far away" from the threshold. In other situations, a polynomial may be a bad approximation, and smaller biases will occur with a local linear regression.

  In practice, parametric and non-parametric approaches lead to the computation of the exact same statistic. For example, the procedure of regressing the outcome Y on V and a treatment dummy D and an interaction $V \cdot D$, can be viewed as a "parametric" regression *or* a local linear regression with

[34] See Lee and Card (2008) for a discussion.

[35] As an example, Powell (1994) points out that the same least squares estimator can simultaneously viewed as solutions to parametric, semi-parametric, and nonparametric problems.

[36] Unless the underlying function is exactly linear in the area being examined.

a very large bandwidth. Similarly, if one wanted to exclude the influence of data points in the tails of the X distribution, one could call the exact same procedure "parametric" after trimming the tails, or "non-parametric" by viewing the restriction in the range of X as a result of using a smaller bandwidth.[37] Our main suggestion in estimation is to not rely on one particular method or specification. In any empirical analysis, results that are stable across alternative and equally plausible specifications are generally viewed as more reliable than those that are sensitive to minor changes in specification. RD is no exception in this regard.

- **Goodness-of-fit and other statistical tests can help rule out overly restrictive specifications.** Often the consequence of trying many different specifications is that it may result in a wide range of estimates. Although there is no simple formula that works in all situations and contexts for weeding out inappropriate specifications, it seems reasonable, at a minimum, not to rely on an estimate resulting from a specification that can be rejected by the data when tested against a strictly more flexible specification. For example, it seems wise to place less confidence in results from a low-order polynomial model, when it is rejected in favor of a less restrictive model (e.g., separate means for each discrete value of V). Similarly, there seems little reason to prefer a specification that uses all the data, if using the same specification but restricting to observations closer to the threshold gives a substantially (and statistically) different answer.

A Recommended "Checklist" for Implementation

Below we summarize the recommendations given by Lee and Lemieux (2009) for the analysis, presentation, and estimation of RD designs.

1. **To assess the possibility of manipulation of the assignment variable, show its distribution.** The most straightforward thing to do is to present a histogram of the assignment variable, after partitioning the support of V into intervals; in practice, V may have a natural discreteness to it. The bin widths should as small as possible, without compromising the ability to visually see the overall shape of the

[37] One of the reasons why typical non-parametric "methods" (e.g. local linear regression) are sometimes viewed as being superior is that the statistics yield consistent estimators. But it is important to remember that such consistency is arising from an asymptotic approximation that dictates that one of the "parameters" of the statistic (i.e. the function of the sample data) – the bandwidth – shrinks (at an appropriate rate) as the sample size increases. Thus, the consistency of the estimator is a direct result of a different notion of asymptotic behavior. If one compares the behavior of "non-parametric" statistics (e.g. local linear regression) with that of "parametric" statistics (e.g. global polynomial regression) using the same asymptotic framework (i.e. statistics are not allowed to change with the sample size), then the non-parametric method loses this superiority in terms of consistency. Depending on the true underlying function (which is unknown), the difference between the truth and the probability limit of the estimator, may be larger or smaller with the "parametric" statistic.

distribution. The bin-to-bin jumps in the frequencies can provide a sense in which any jump at the threshold is "unusual". For this reason, we recommend *against* plotting a smooth function comprised of kernel density estimates. A more formal test of a discontinuity in the density can be found in McCrary (2008).

2. **Present the main RD graph using binned local averages.** As with the histogram, the recommendation is to graphically present the sample means within the defined intervals. The non-overlapping nature of the bins for the local averages is important; we recommend against simply presenting a continuum of nonparametric estimates (with a single break at the threshold), as this will naturally tend to give the impression of a discontinuity even if there does not exist one in the population. Lee and Lemieux (2009) suggest a cross-validation procedure as well as simple ways to test the bin width choice against less restrictive alternatives. They recommend generally "undersmoothing", while at the same time avoiding "too narrow" bins that produce a scatter of data points, from which it is difficult to see the shape of the underlying function. Indeed, they also recommend against simply plotting the raw data without a minimal amount of local averaging.

3. **Graph a benchmark polynomial specification.** Super-impose onto the above graph the predicted values from a low-order polynomial specification. One can often informally assess, by comparing the two functions, whether a simple polynomial specification is an adequate summary of the data. In a way, these two functions give a sense of the range of functions that would fit the data. On the one hand, the local averages represent a flexible "non-parametric" representation of the true underlying function. On the other hand, a polynomial represents a "best case" scenario in terms of the variance of the RD estimate: if the true function really is a polynomial of the chosen order, standard regression theory suggests that the least squares estimator (that uses all the observations) will be unbiased, and potentially efficient in a class of all linear unbiased estimators.

4. **Explore the sensitivity of the results to a range of bandwidths, and a range of orders to the polynomial.Lee and Lemieux (2009)** provide an example of how to systematically examine different degrees of smoothing, through different bandwidths or polynomial orders, using both cross-validation to provide a rough guide to sensible bandwidths, and the Akaike Information Criterion (AIC) as a rough guide to sensible orders for the polynomial. A useful graphical device for illustrating the sensitivity of the results to bandwidths is to plot the local linear discontinuity estimate against a continuum

of bandwidths (within a range of bandwidths that are not ruled out by available specification tests). For an example of such a presentation, see the online appendix to Card et al. (2009a) and Lee and Lemieux (2009).

5. **Conduct a parallel RD analysis on the baseline covariates.** As discussed earlier, if the assumption that there is no precise manipulation or sorting of the assignment variable is valid, then there should be no discontinuities in variables that are determined prior to the assignment.

6. **Explore the sensitivity of the results to the inclusion of baseline covariates.** As discussed above, in a neighborhood of the discontinuity threshold, pre-determined covariates will have the same distribution on either side of the threshold, implying that inclusion of those covariates in a local regression should not affect the estimated discontinuity. If the estimates do change in an important way, it may indicate a potential sorting of the assignment variable that may be reflected in a discontinuity in one or more of the baseline covariates. Lee and Lemieux (2009) show how the assumption that the covariates can be approximated by the same order of polynomial as Y as a function of V can be used to justify including the covariates linearly in a polynomial regression.

Although it is impractical for researchers to present every permutation of presentation (e.g. points 2-4 for every one of 20 baseline covariates), probing the sensitivity of the results to this array of tests and alternative specifications – even if they only appear in online appendices – is an important component of a thorough RD analysis.

3.4 Regression Discontinuity Design: Fuzzy

Returning to our hypothetical job search assistance program, consider the same setup as the RD design described above: based on past employment and earnings and a model of who would most benefit from the program, the government constructs a score V where those with $V \geq 0$ will receive the treatment (phone call/personal visit of job counselor). But now assume that V crossing the threshold 0 only determines whether the agency explicitly provides information to the individual about the existence of the program. In other words, V determines Z – as defined in the case of random assignment with imperfect compliance, discussed above in Section3.2. As a result, D is no longer a deterministic function of V, but $\Pr[D = 1|V]$ is a potentially discontinuous function in V. This is known as the "fuzzy" RD design. (See Hahn et al. (2001), for example, for a formal definition).

The easiest way to understand the fuzzy RD design is to keep in mind that the relationship between the fuzzy design and the sharp design parallels the relation between random assignment with imperfect compliance and random assignment with perfect compliance. Because of this parallel, our assessment of the potential internal validity of the design and potential caveats follows the discussion in Section 3.2. Testing the design follows the same principles as in Section 3.2. Furthermore, one could, in principle, combine the extrapolative ideas in Section 3.3.2 and Section 3.2.2 to use fuzzy RD design estimates to make extrapolations along two dimensions: predicting the effect under mandatory compliance, and predicting the average effects at points away from the threshold.

We limit our discussion here to making explicit the conditions for identification of average effects within the common econometric framework we have been utilizing. The different conditions for this case are

- D6: (Discontinuous rule for Z) $Z = 1[V \geq 0]$. This implies that P^* in Equation (2) is given by the function $p^*(W,U) = \Pr[V \geq 0|W,U]\, p_1^*(W,U) + \Pr[V < 0|W,U]\, p_0^*(W,U)$, where $p_z^*(W,U) \equiv \Pr[D = 1|W,U,Z = z]$.

- S14: (Exclusion Restriction) $Y = y(D,W,U,V)$, where $y(d,x,w,v)$ is continuous in v (at least in a neighborhood of $v = 0$). $W = w(U,V)$, where $w(u,v)$ is continuous in v (at least in a neighborhood of $v = 0$). (Z does not enter either function).

In sum, we have 1) the conditions from the Sharp RD (D5 (V observable), S11 (Positive Density at Threshold), S13 (Continuous Density at Threshold)), 2) a condition from the randomized experiment with imperfect compliance (S9 (Probabilistic Monotonicity)), and 3) two new "hybrid" conditions – D6 (Discontinuous rule for Z) and S14 (Exclusion Restriction).

Given the discontinuous rule D6, we have

$$\begin{aligned} E[Y|V = 0] &= \int y(0, w(u), u, 0) + p_1^*(w(u), u)\, \Delta(w(u), u, 0)\, dF_{U|V=0}(u) \\ E[Y|V = v] &= \int y(0, w(u), u, v) + p_0^*(w(u), u)\, \Delta(w(u), u, v)\, dF_{U|V=v}(u) \end{aligned}$$

and because of S11, S13, and S14, we can combine the integrals so that the difference equals

$$E[Y|V = 0] - \lim_{v \uparrow 0} E[Y|V = v] = \int \Delta(w(u), u, 0)\, (p_1^*(w(u), u) - p_0^*(w(u), u))\, \frac{f_{V|U=u}(0)}{f_V(0)}\, dF_U(u)$$

Normalizing the difference by the quantity $E[D|V=0]-\lim_{v\uparrow 0}E[D|V=v]$ yields

$$\frac{E[Y|V=0]-\lim_{v\uparrow 0}E[Y|V=v]}{E[D|V=0]-\lim_{v\uparrow 0}E[D|V=v]}=\int\Delta(w(u),u,0)\left[\frac{(p_1^*(w(u),u)-p_0^*(w(u),u))}{E[D|V=0]-\lim_{v\uparrow 0}E[D|V=v]}\frac{f_{V|U=u}(0)}{f_V(0)}\right]dF_U(u)$$

where the normalizing factor ensures that the weights in square brackets average to one.[38]

Thus the fuzzy RD estimand is a weighted average of treatment effects. The weights reflect two factors: the relative likelihood that a given type U's V will be close to the threshold reflected in the term $\frac{f_{V|U=u}(0)}{f_V(0)}$, and the influence that V crossing the threshold has on the probability of treatment, as reflected in $p_1^*(w(u),u)-p_0^*(w(u),u)$. S9 ensures these weights are nonnegative.

From a purely ex ante evaluation perspective, the weights would seem peculiar, and not related to any meaningful economic concept. But from a purely ex post evaluation perspective, the weights are a statement of fact. As soon as one believes that there is causal information in comparing Y just above and below the threshold – and such an intuition is entirely driven by the institutional knowledge given by D5 and D6 – then it appears that the *only* way to obtain *some* kind of average effect (while remaining as agnostic as possible about the other unobservable mechanisms that enter the latent propensity P^*) with the data Y,V,D, is to make sure that the implied weights integrate to 1. We have no choice but to divide the difference by $E[D|V=0]-\lim_{v\uparrow 0}E[D|V=v]$.

As we have argued in previous sections, rather than abandon potentially highly credible causal evidence because the data and circumstance that we are handed did not deliver us the "desired" weights, we believe a constructive approach might leverage off the credibility of quasi-experimental estimates, and use them as inputs in an extrapolative exercise that will necessarily involve imposing more structure on the problem.

4 Research Designs Dominated by Self-Selection

In this section, we briefly consider a group of research designs where institutional knowledge of the treatment assignment process typically does not provide most of the information needed to draw causal inferences. In Section 3, we discussed how some aspects of the assignment process could be treated more as literal "descriptions" ("D" conditions), rather than conjectures or assumptions. In each of the four research

[38] Note that $\int\left[\left(p_1^*(w(u),u)-p_0^*(w(u),u)\right)\frac{f_{V|U=u}(0)}{f_V(0)}\right]dF_U(u)=\int p_1^*(w(u),u)\frac{f_{V|U=u}(0)}{f_V(0)}dF_U(u)-\lim_{v\uparrow 0}\int p_0^*(w(u),u)\frac{f_{V|U=u}(v)}{f_V(v)}dF_U(u)=\int p_1^*(w(u),u)\,dF_{U|V=0}(u)-\lim_{v\uparrow 0}\int p_0^*(w(u),u)\,dF_{U|V=v}(u)=E[D|V=0]-\lim_{v\uparrow 0}E[D|V=v]$.

designs, those "D" conditions went a long way towards identification, and when other structural assumptions ("S" conditions) were needed, the class of models consistent with those assumptions, while strictly smaller because of the restrictions, arguably remained very broad.

With the common program evaluation approaches considered in this section, we shall see that the assignment process is not dominated by explicit institutional knowledge, and identification thus requires more conjectures/assumptions ("S" conditions) to make causal inferences. We will argue that with these designs there will be more scope for alternative plausible economic models that would be strictly inconsistent with the conditions needed for identification. Of the three approaches we consider, the "difference-in-difference" approach appears to have the best potential for testing the key "S" conditions needed for identification.

For the reasons above, we suggest that these designs will tend to deliver causal inferences with lower internal validity, in comparison to the designs described in Section 3. But even if one agrees with the three criteria we have put forth in Section 2.2.1 to assess internal validity, and even if one agrees with the conclusion that these designs deliver lower internal validity, the question of "how much lower" requires a subjective judgment, and such a question is ill-defined at any rate (what is a unit of "internal validity"?). Of the three criteria we discuss, the extent to which the conditions for identification can be treated as a hypothesis with testable implications seems to be the least subjective in nature.

In our discussion below, it is still true that a particular weighted average effect of "interest" from an ex ante evaluation problem may in general be quite different from the effects identified by these research designs. In this sense, these approaches suffer from similar "external validity" concerns as discussed in Section 3. We will therefore focus our discussion on the ex post evaluation goal, and do not have separate sections on ways to extrapolate from the results of an ex post evaluation to forecast the effect of interest in an ex ante evaluation.

4.1 Using Longitudinal Data: Difference-in-Difference

We now consider the case where one has longitudinal data on program participants and non-participants. Suppose we are interested in the effectiveness of a job training program in raising earnings. Y is now earnings and D is participation in the job training program. A commonly used approach in program evaluation is the "difference-in-difference" design, which has been discussed as a methodology and utilized in program evaluation research in some form or another countless times. Our only purpose here is to discuss how the design fits into the general framework we have used in this chapter, and to be explicit about the restrictions

in a way that facilitates comparison with the designs in Section 3.

First, let us simplify the problem by considering the situation where the program was made available at only one point in time τ. This allows us to define $D = 1$ as those who were treated at time τ, and $D = 0$ as those who did not take up the program at that time.

- D7 (Program exposure at one point in time): Individuals will have $D = 0$ for all $t < \tau$; for $t \geq \tau$, the non-treated will continue to have $D = 0$ while the treated will have $D = 1$.

W will continue to denote all the factors that could potentially affect Y. Additionally, we imagine that this vector of variables could be partitioned into sub-vectors W_t, $t = 1, \ldots, \tau$. where the subscript denotes the value of the variables at time t.

Furthermore, we explicitly include time in the outcome equation as

$$Y_t = y(D, W, U, t)$$

A difference-in-difference approach begins by putting some structure on Y:

- S15 (Additive Separability): $y(D, W, U, t) = g(D, W, U) + \alpha(W, U) + h(W, t)$.

This highly restrictive structure (although it does not rule out heterogeneous treatment effects) is the standard "individual fixed effects" specification for the outcome, where $\alpha(W, U)$ captures the permanent component of the outcome.

Perhaps the most important thing to keep in mind is that D7 and S15 is *not* generally sufficient for the difference-in-difference approach to identify the treatment effects. This is because the two differences in question are

$$\begin{aligned} E[Y_\tau - Y_{\tau-k} | D = 1] &= E[g(1, W, U) - g(0, W, U) | D = 1] + \int h(W, \tau) - h(W, \tau - k) \, dF_{U|D=1}(u) \\ E[Y_\tau - Y_{\tau-k} | D = 0] &= \int h(W, \tau) - h(W, \tau - k) \, dF_{U|D=0}(u) \end{aligned}$$

The term $E[g(1, W, U) - g(0, W, U) | D = 1]$ is equal to $E[y(1, W, U, \tau) - y(0, W, U, \tau) | D = 1]$, the treatment on the treated (TOT) parameter. When second equation is subtracted from the first, the terms with h do not cancel without further restrictions.

One approach to this problem is to further assume that

- S16 (Influence of "Other factors" Fixed): $h(W,t) - h(W,t-k) = \gamma(t,t-k)$, for any k.

This certainly would ensure that the D-in-D estimand identifies the TOT. It is, however, restrictive: even if $h(W,t) = h(W_t)$ – so that only contemporaneous factors are relevant – then as long as there were some factors in W that changed over time, this would be violated. Note that in this case, it is irrelevant how similar or different the distribution of unobservable types are between the treated and non-treated individuals ($F_{U|D=1}(u)$ vs. $F_{U|D=0}(u)$).[39]

4.1.1 Assessment

In terms of our three criterion for assessing this approach, how does the D-in-D fare? It should be very clear from the above derivation that the model of both the outcome and treatment assignment is far cry from a literal description of the data generating process, except for D7, which describes the timing of the program and structure of the data. As for the second criterion, it is not difficult to imagine writing down economic models that would violate the restrictions. Indeed, much of the early program evaluation literature (Heckman and Robb Jr., 1985; Ashenfelter, 1978; Ashenfelter and Card, 1985) discussed different scenarios under which S15 and S16 would be violated, and how to nevertheless identify program effects.

On the other hand, there is one positive feature of this approach – and it is driven by S16, which is precisely the assumption that allowed identification of the program effect – is that there are strong testable predictions of the design, namely

$$E[Y_\tau - Y_{\tau-k}|D=1] - E[Y_\tau - Y_{\tau-k}|D=0] = E[g(1,W,U) - g(0,W,U)|D=1]$$

for all $k > 0$. That is, the choice of the base year in constructing the DD should be irrelevant. Put differently, it means that

$$E\left[Y_{\tau-k} - Y_{\tau-k-j}|D=1\right] - E\left[Y_{\tau-k} - Y_{\tau-k-j}|D=0\right] = 0$$

for $j,k > 0$: the DD estimand during the "pre-program" period should equal zero. As is well-known from the literature, there are as many testable restrictions as there are pre-program periods, and it is intuitive that the more evidence that these restrictions are not rejected, the greater confidence we might have in the causal inferences that are made from the D-in-D.

[39] S16 has the implication that there will be no variance in $Y_\tau - Y_{\tau-k}$ for the untreated group, which will in practice almost never be the case. The variance in changes could be accommodated by an independent, additive error term in the outcome equation.

Overall, while it is clear that the assumptions given by S15 and S16 require a great deal of speculation – and arguably a greater suspension of disbelief, relative to the conditions outlined in Section 3– at least there is empirical evidence (the pre-program data) that can be used to assess the plausibility of the key identifying assumption, S16.

4.2 Selection on Unobservables and Instrumental Variables

In this section, we briefly discuss the identification of program effects when we have much less information about treatment assignment, relative to the designs in Section 4. We will focus on the use of instrumental variable approaches, but it will be clear that the key points we discuss will equally apply to a control function approach.

The instrumental variable approach is typically described as finding a variable Z that impacts Y only through its effect on treatment status D. Returning to our job search assistance program example, let us take Z to be the binary variable of whether the individual's sibling participated in the program. The "story" behind this instrument would be that the sibling's participation might be correlated with the individual's participation – perhaps because the sibling would be an influential source of information about the program – but that there is no reason why a sibling's participation in the program would directly impact the individual's re-employment probabilities.

Even if one completely accepts this "story" – and there are undoubtedly reasons to question its plausibility – this is not sufficient to identify the treatment effect via this instrument. To see this, we use the framework in Equations (1), (2) and (3), and adopt S8 (Excludability) and S9 (Probabilistic Monotonicity).

S8 is the formal way to say that for any individual type U, Z (the sibling's program participation) has no indirect nor direct impact on the outcome. This exclusion restriction might come from a particular behavioral model. Furthermore, S9 simply formalizes the notion that for any given type U, the probability of receiving treatment is higher than if a sibling participated in the program. Alternatively, it says that for each type U, there are more individuals who are induced to receive treatment because of their sibling's participation, than those who would be discouraged from doing so.

The problem is that, in general, it is easy to imagine that there is heterogeneity in the latent propensity for the individual to have a sibling participate in the program: $P_Z^* = \Pr[Z = 1|U]$ has a non-degenerate distribution. If such variability in P_Z^* exists in the population, this immediately implies that $F_{U|Z=1}(u)$ will in general be different from $F_{U|Z=0}(u)$. The IV (Wald) estimand will in general not identify any average

effect, LATE or otherwise.

Typically researchers immediately recognize that their instrument Z is not "as good as randomly assigned" as in D3, and so instead appeal to a "weaker" condition that

- S17 (Conditional on W, Z "as good as randomly assigned"): $\Pr[Z=1|U=u]=p_{z1}(w(u))$, a function of w.

This is a restriction on the heterogeneity in P_Z^*. S17 says that types with the same W have identical propensities P_Z^*.

Of course, the notion that the analyst knows and could measure all the factors in W, is a conjecture in itself:

- S18 (Sufficient variables for P_Z^*): Let $X=x(U)$ be the observable (to the researcher) elements of W, and assume $p_{z1}(w(u))=p_{z1x}(x(u))$ for all u.

S18 simply says that the researcher happens to observe all the variables that determine the propensity P_Z^*.

It should be clear that with S17, S18, S8, and S9, one can condition the analysis on a particular value $X=x$, and apply the results from the randomized experiment with imperfect compliance (Section3.2).

Note that while we have focused on the binary instrument case, it should also be clear that this argument will apply to the case when Z is continuously distributed. It is not sufficient for Z to simply be excluded from the outcome equation, Z must be assumed to be (conditionally) independent of U, and indeed this is the standard assumption in the evaluation literature [40]

4.2.1 Assessment

It is clear that in this situation, as D3 is replaced with S17 and S18, there is now *no* element of the statistical model that can be considered a literal description of the treatment assignment process. The model is entirely driven by assumptions about economic behavior, rather than a description of the data generating process that is derived from our institutional knowledge about how treatment was assigned.

S17 makes it clear that causal inferences will be dependent on having the correct set of variables X. Without the complete set of X, there will be variability in P_Z^* conditional on the covariates, which will mean that the distribution of types U will not be the same in the $Z=1$ and $Z=0$ groups. In general,

[40] In the more general discussion in Heckman and Vytlacil (2005), for example, Z is assumed, at a minimum, to be independent of U_1 and U_0 (or of potential outcomes) given a set of observed "conditioning variables."

different theories about which X's satisfy S18 will lead to a different causal inference. Recall that no similar specification of the relevant X's was necessary in the case of the randomized experiment with imperfect compliance, considered in Section 3.2.

Finally, there seems to be very little scope for testing the validity of this design. If the argument is that the instrument is independent of U only conditional on all the Xs observed by the researcher, then all the data will have been "used up" to identify the causal parameter.

To make the design somewhat testable, the researcher could assume that only a smaller subset of variables in X are needed to characterize the heterogeneity in P_Z^*. In that case, one could imagine conditioning on that smaller subset, and examining whether the distribution of the remaining X variables are balanced between the $Z = 1$ and $Z = 0$, as suggested in Section 3.2 (with the appropriate caveats and qualifications discussed there). But in practice, when evidence of imbalance is found, the temptation is to simply include those variables as conditioning variables to achieve identification, which then eliminates the potential for testing.

What this shows is the benefit to credibility that one obtains from explicit knowledge about the assignment process whereby D3 is a literal description of what is known about the process. It disciplines the analysis so that any observed X variables can be used to treat D3 as a hypothesis to be tested.

4.3 Selection on Observables and Matching

Absent detailed institutional knowledge of the the selection process, i.e., the propensity score equation, a common approach to the evaluation problem is to "control" for covariates, either in a multiple regression framework, or more flexibly through multivariate matching or propensity score techniques.[41] Each of these approaches essentially assumes that conditional on some observed covariates X, treatment status is essentially "as good as randomly assigned".

In terms of the model given in Equations (1), (2), and (3), one can think of the selection on observables approach as amounting from two important assumptions. First we have

- S19 (Conditional on W, D is "as good as randomly assigned"): $P^* = p^*\left(w(U)\right)$, a function of w.

Here, the unobservable type U does not enter the latent propensity, whereas it does in (2).

[41] For a discussion of several of these approaches, see Busso et al. (2008) and Busso et al. (2009).

It is important to reiterate that the function $\Pr[D=1|W=w]$, the so-called "propensity score" – which can be obtained as long as one can observe W – is *not,* in general, the same thing as the latent propensity P^* for a given $W=w$. That is, even though it will always be true that $\Pr[D=1|W=w]=E[P^*|W=w]$, there will be heterogeneity in P^* for a given $W=w$, unless one imposes the condition S19. And it is precisely this heterogeneity, and its correlation with the outcome Y, that is the central problem of making causal inferences, as discussed in Section 2.2.

In addition, if the researcher presumes that there are some factors that are unobservable in W, then one must further assume that

- S20 (S19 + Sufficient variables for P^*): Let $X=x(U)$ be the observable (to the researcher) elements of W, and assume $p^*(w(u))=p^*_x(x(u))$ for all u.

So S20 goes further to say that not only are the Xs sufficient to characterize the underlying propensity, all the unobservable elements of W are irrelevant in determining the underlying propensity P^*.

S20 has the same implications as condition D2 discussed in Section 3.1.2: the difference $E[Y|D=1,X=x]-E[Y|D=0,X=x]$ identifies the (conditional) average treatment effect $E[\Delta(W,U)|X=x]$. The key difference is that in Section 3.1.2, D2 was a literal description of a particular assignment process (random assignment with probabilities of assignment being a function of X).[42] Here, S20 is a restriction on the framework defined by Equations (1), (2), and (3).

To see how important it is not to have variability in P^* conditional on X, consider the "conditional" version of Equation (5)

$$\begin{aligned} E[Y|D=1,X=x]-E[Y|D=0,X=x] &= \int E[y(1,w(U),U)|P^*=p^*,X=x]\, f_{P^*|D=1,X=x}(p^*)\,dp^* \\ &\quad - \int E[y(0,w(U),U)|P^*=p^*,X=x]\, f_{P^*|D=0,X=x}(p^*)\,dp^* \end{aligned}$$

A non-degenerate density $f_{P^*|X=x}(p^*)$ will automatically lead to $f_{P^*|D=1,X=x}(p^*)\,dp^* \neq f_{P^*|D=0,X=x}(p^*)\,dp^*$, which would prevent the two terms from being combined.[43]

[42] Indeed, for a "half century" the basic framework was "entirely tied to randomization based evaluations" and was "not perceived as being relevant for defining causal effects in observational studies."Rubin (1990)

[43] Recall that $f_{P^*|D=1,X=x}(p^*)\,dp^* = \frac{p^*}{\Pr[D=1|X=x]} f_{P^*|X=x}(p^*)\,dp^*$ and $f_{P^*|D=0,X=x}(p^*)\,dp^* = \frac{1-p^*}{\Pr[D=0|X=x]} f_{P^*|X=x}(p^*)\,dp^*$, which will be unequal with $f_{P^*}(p^*)$ non-degenerate.

4.3.1 Assessment: Included Variable Bias

In most observational studies, analysts will rarely claim that they have a model of behavior or institutions that dictate that the assignment mechanism *must* be modeled as S20. More often, S20 is invoked because there is an explicit recognition that there is non-random selection into treatment, so that D is certainly not unconditionally randomly assigned. S20 is offered as a "weaker" alternative.

Perhaps the most unattractive feature of this design is that, even if one believes that S20 does hold, typically there is not much in the way of guidance as to what Xs to include, as has long been recognized (Heckman et al., 1998a). There usually is a multitude of different plausible specifications, and no disciplined way to choose among those specifications.

It is therefore tempting to believe that if we compare treatment and control individuals who look more and more similar on observable dimensions X, then – even if the resulting bias is non-zero – at the least, the bias in the estimate will decrease. Indeed, there is a "folklore" in the literature which suggests that "overfitting" is either beneficial or at worst harmless. Rubin. and Thomas (1996) suggest including variables in the propensity score unless there is a consensus that they do not belong. Millimet and Tchernis (2009) go further and argue that *overfitting* the propensity score equation is possibly beneficial and at worst harmless.

It is instructive to consider a few examples to illuminate why this is in general not true, and how adding X's can lead to "included variable bias". To gain some intuition, first consider the simple linear model

$$Y = \beta_0 + D\beta_1 + \varepsilon$$

where β_1 is the coefficient of interest and $COV(D,\varepsilon) \neq 0$. The probability limit of the OLS regression coefficient on D is

$$\beta_{OLS} = \beta_1 + \frac{COV(D,\varepsilon)}{VAR(D)}$$

Now suppose there is a "control variable" X. Suppose X actually has covariance $COV(X,\varepsilon) = 0$, so it is an "irrelevant" variable, but it can explain some variation in D. When X is included, the least squares coefficient on D will be

$$\beta_{OLS,X} = \frac{COV\left(Y, D-\hat{D}\right)}{VAR\left(D-\hat{D}\right)} = \beta_1 + \frac{COV(D,\varepsilon)}{VAR\left(D-\hat{D}\right)}$$

where $\hat{D}$ is the predicted value from a population regression of D on X. This expression shows that the magnitude of the bias in the least squares estimand that includes X will be strictly larger, with the denominator

in the bias term decreasing. What is happening is that the extra variable X is doing nothing to reduce bias, while absorbing some of the variation in D.

To gain further intuition on the potential harm in "matching on Xs" in the treatment evaluation problem, consider the following system of equations

$$\begin{aligned} Y &= \beta_0 + D\beta_1 + X\beta_2 + U \\ D &= 1\left[\delta_0 + X\delta_1 + V \geq 0\right] \end{aligned}$$

where X, the "control" variable, is in this case a binary variable. U, V is assumed to be independent of X. This is a simplified linear and parametric version of (1), (2), and (3).[44]

The bias of the simple difference in means – without accounting for X – can be shown to be

$$\begin{aligned} BIAS_{DIF} &\equiv E\left[Y|D=1\right] - E\left[Y|D=0\right] - \beta_1 \\ &= \beta_2\left(\Pr\left[X=1|D=1\right] - \Pr\left[X=1|D=0\right]\right) + \left\{E\left[U|D=1\right] - E\left[U|D=0\right]\right\} \end{aligned}$$

whereas the bias in the matching estimand for the TOT is

$$\begin{aligned} BIAS_{MATCH} &\equiv \left[\sum_x \Pr\left[X=x|D=1\right]\left(E\left[Y|D=1,X=x\right] - E\left[Y|D=0,X=x\right]\right)\right] - \beta_1 \\ &= \left[\sum_x \Pr\left[X=x|D=1\right]\left(E\left[U|D=1,X=x\right] - E\left[U|D=0,X=x\right]\right)\right] \\ &= E\left[U|D=1\right] - \left(E\left[U|D=0,X=0\right]\Pr\left[X=0|D=1\right] + E\left[U|D=0,X=1\right]\Pr\left[X=1|D=1\right]\right) \end{aligned}$$

In this very simple example, a comparison between $BIAS_{DIF}$ and $BIAS_{MATCH}$ reveals two sources of differences. First, there is a standard "omitted variable bias" that stems from the first term in $BIAS_{DIF}$. This bias does not exist in the matching estimand.

But there is another component in $BIAS_{DIF}$, the term in curly braces – call it the "selectivity bias" term. This term *will always be smaller in magnitude than* $|BIAS_{MATCH}|$. That is, "controlling for" X *can only increase* the magnitude of the selectivity bias term. To see this, note that the difference between the term in curly braces in $BIAS_{DIF}$ and $BIAS_{MATCH}$ is the difference between

$$E\left[U|D=0\right] = E\left[U|D=0,X=0\right]\Pr\left[X=0|D=0\right] + E\left[U|D=0,X=1\right]\Pr\left[X=1|D=0\right] \qquad (15)$$

[44] Here, $P^* = \Pr\left[V \geq -\delta_0 - X\delta_1|X,U\right]$.

and

$$E[U|D=0,X=0]\Pr[X=0|D=1]+E[U|D=0,X=1]\Pr[X=1|D=1] \tag{16}$$

in $BIAS_{MATCH}$. Each of these expressions is a weighted average of $E[U|D=0,X=x]$.

Consider the case of positive selectivity, so $E[U|V=v]$ is increasing in v, so that the selectivity term (curly braces) in $BIAS_{DIF}$ is positive, and suppose that $\Pr[D=1|X=1] > \Pr[D=1|X=0]$ (i.e. $\delta_1 > 0$).[45] This means that $E[U|D=0,X=1] = E[U|V<-\delta_0-\delta_1] < E[U|V<-\delta_0] = E[U|D=0,X=0]$. That is, among the non-treated group, those with $X=1$ are *more* negatively selected than those with $X=0$.

Comparing (15) and (16), it is clear that $BIAS_{MATCH}$ automatically places relatively more weight on $E[U|D=0,X=1]$ – since $\Pr[X=1|D=1] > \Pr[X=1|D=0]$ – , and hence $BIAS_{MATCH}$ will exceed the selectivity term (curly braces) in $BIAS_{DIF}$.[46] Intuitively, as X can "explain" more and more of the variation in D, the exceptions (those with $X=1$, but $D=0$) must have unobservable factors that are even more extreme in order to be exceptional. And it is precisely those exceptional individuals that are implicitly given relatively more weight when we "control" for X.

So in the presence of nontrivial selection on unobservables, a matching on observables approach will generally *exacerbate* the selectivity bias. Overall, this implies that a reduction in bias will require the possible "benefits" – elimination of the omitted variable bias driven by β_2 – to outweigh the cost of exacerbating the selectivity bias.

There is another distinct reason why the magnitude of $BIAS_{MATCH}$ may be larger than that of $BIAS_{DIF}$. The problem is that β_2 is unknown, and the sign and magnitude need not be tied to the fact that U correlates with V. In the above example, even if there is positive selectivity on unobservables, β_2 may well be negative, and therefore, $BIAS_{DIF}$ could be zero (or very small). So even if matching on X had a small effect on the selectivity bias component, the elimination of the omitted variable bias term will cause $BIAS_{MATCH} > BIAS_{DIF}$. That is, if the two sources of biases were offsetting each other in the simple difference, eliminating one of the problems via matching make the overall bias increase.

Overall, we conclude that as soon as the researcher admits departures from S20, there is a rather weak case to be made for "matching" on observables being an improvement. Indeed, there is a compelling argument that including more X's will increase bias, and that the "cure may be worse than the disease".

Finally, in terms of the third criterion we have been considering, the matching approach seems to have

[45] Parallel arguments hold when $E[U|V=v]$ is decreasing in v and/or when $\delta_1 < 0$.

[46] By Bayes' rule, $\Pr[X=1|D=1] = \frac{\Pr[D=1|X=1]}{\Pr[D=1]}\Pr[X=1] > \frac{1-\Pr[D=1|X=1]}{1-\Pr[D=1]}\Pr[X=1] = \Pr[X=1|D=0]$.

no testable implications whatsoever. One possibility is to specify a particular subset X' of the Xs that are available to the researcher, and make the argument that it is specifically those variables that determine treatment in S20. The remainder of the observed variables could be used to test the implication that the distribution of types U is the same between the the treated and non-treated populations, conditional on X'. The problem, of course, is that if some differences were found in those Xs not in X', there would again be the temptation to simply include those variables in the subset X'. Overall, not only do we believe this design to have a poor theoretical justification in most contexts (outside of actual stratified randomized experiments), but there seems to be nothing in the design to discipline which Xs to include in the analysis, and as we have shown above, there is a great risk to simply adding as many Xs to the analysis as possible.

4.3.2 Propensity Score, Matching, Re-weighting: Methods for Descriptive, Non-Causal Inference

Although we have argued that the matching approach is not compelling as a research design for causal inference, it can nevertheless be a useful tool for descriptive purposes. Returning to our hypothetical job search assistance program, suppose that the program is voluntary, and that none of the data generating processes described in Section 4 apply. We might observe the difference

$$E[Y|D=1]-E[Y|D=0]$$

but also notice that the distribution of particular Xs (education, age, gender, previous employment history) are also different: $F_{X|D=1}(x) \neq F_{X|D=0}(x)$. One could ask the descriptive question, "mechanically, how much of the difference could be *exclusively* explained by differences in the distribution of X?" We emphasize the word "mechanically"; if we observe that Y varies systematically by different values of X for the treated population, and if we further know that the distribution of X is different in the non-treated population, then even if the program were entirely irrelevant, we would nevertheless generally *expect* to see a difference between $E[Y|D=1]$ and $E[Y|D=0]$.

Suppose we computed

$$E\left[\widehat{Y|D=0}\right] \equiv \int E[Y|D=0,X=x]\, dF_{X|D=1}(x)$$

Then the difference

$$E[Y|D=1] - E\left[\widehat{Y|D=0}\right] \tag{17}$$

would tell us how relevant D is in predicting Y, once adjusting for the observables X. If one adopted S20 then this could be interpreted as an average treatment effect for the treated. But more generally, this adjusted difference could be viewed as a descriptive, summary statistic, in the same way multiple regressions could similarly provide descriptive information about the association between Y and D after partialling out X.

We only briefly review some of the methods used to estimate the quantity (17), since this empirical exercise is one of the goals of more general decomposition methods, which is the focus of the chapter by Firpo et al. (2010). We refer the reader to that chapter for further details.

Imputation: Blinder/Oaxaca

One way of obtaining (17), is to take each individual in the treated sample, and "impute" the missing quantity, the average Y given the individual's characteristics X. This is motivated by the fact that

$$\int\int [y - E[Y|D=0, X=x]] f_{X,Y|D=1}(x,y)\, dxdy \tag{18}$$

is identical to the quantity (17).

The sample analogue is given by

$$\frac{1}{N_1} \sum_{i:D_i=1} [Y_i - \widehat{Y}_i]$$

where N_1 is the number of observations in the treated sample, and $\hat{Y}_i$ is the predicted value of regressing Y on X for the non-treated sample. This is immediately recognizable as a standard Blinder/Oaxaca exercise.

Matching

One development in the recent labor economics literature is an increased use of matching estimators and estimators based on the propensity score and semi-parametric estimators which eschew parametric specification of the outcome functions. The concern is that the regression used to predict $\hat{Y}_i$ may be a bad approximation of the true conditional expectation.

The first approach is to simply use the sample mean of Y_i for all individuals in the non-treated sample that have exactly the same value for X as the individual i . Sometimes it will be possible to do this for every individual (e.g. when X is discrete and for each value of X there are treated and non-treated observations).

In other cases, X is so multi-dimensional that for each value of X there are very few observations with

many values only having treated or non-treated observations. Alternatively, X could have continuously distributed elements, in which case exact matching is impossible. In this case, one approach is to compute non-parametric estimates of $\hat{Y}_i$ using kernel regression or local polynomial regression (Hahn, 1998; Hirano et al., 2003). A version of matching takes the data point in the control sample that is "closest" to the individual i in terms of the characteristics X, and assigns $\hat{Y}_i$ to be the value of Y for that "nearest match".

Propensity Score Matching

A variant of the above matching approach is to "match" on the Propensity score, rather than on the observed X, and it is motivated by the fact that (17) is also equivalent to

$$\int\int \left[y - E\left[Y|D=0, PS(x)=p\right]\right] f_{PS,Y|D=1}(p,y)\, dpdy$$

where

$$PS(x) = \Pr\left[D=1|X=x\right]$$

is the well-known "propensity score" of Rosenbaum and Rubin (1983). We emphasize once again that PS is *not* the same thing as P^*, the latent propensity to be treated. Indeed it is the fact that there may be variability in P^* conditional on PS, which threatens the validity of the "selection on observables" approach to causal inference.

Re-weighting

An alternative approach is to "re-weight" the control sample so that the re-weighted distribution of X matches that in the treated population. It is motivated by the fact that (18) is also equivalent to

$$E\left[Y|D=1\right] - \int \left(\int y f_{Y|X=x,D=0}(y)\, dy\right) f_{X|D=1}(x)\, dx$$

which is equal to

$$E\left[Y|D=1\right] - \int\int y f_{Y|X=x,D=0}(y) \frac{f_{X|D=1}(x)}{f_{X,Y|D=0}(x,y)} f_{X,Y|D=0}(x,y)\, dxdy$$

Using the fact that $f_{Y|X=x,D=0}(y) = \frac{f_{X,Y|D=0}(x,y)}{f_{X|D=0}(x)}$, this becomes

$$E\left[Y|D=1\right] - \int\int y \left(\frac{f_{X|D=1}(x)}{f_{X|D=0}(x)}\right) f_{X,Y|D=0}(x,y)\, dxdy$$

. The second term is simply a weighted average of Y for the non-treated observations using $\left(\frac{f_{X|D=1}(x)}{f_{X|D=0}(x)}\right)$ as a weight. It is clear that this average will up-weight those individuals with $X = x$, when there is relatively "more" individuals with that value are among the treated than among the non-treated; when there are disproportionately fewer individuals with $X = x$, the weighted average will down-weight the observation.

By Bayes' rule, this weight is also equal to

$$\frac{\Pr[D=1|X=x]}{1-\Pr[D=1|X=x]}\frac{1-\Pr[D=1]}{\Pr[D=1]} = \frac{PS(x)}{1-PS(x)}\frac{1-\Pr[D=1]}{\Pr[D=1]}$$

DiNardo et al. (1996); Firpo (2007)

Thus, in practice, a re-weighting approach will involve computing the sample analogue

$$\frac{1}{N_1}\sum_{i:D_i=1} Y_i - \frac{1}{N_0}\sum_{i:D_i=0}\frac{\hat{PS}(X_i)}{1-\hat{PS}(X_i)}\frac{N_0}{N_1}Y_i$$

where $\hat{PS}(x)$ is the estimated propensity score function for an individual i with $X_i = x$.

A useful aspect of viewing the adjustment as a re-weighting problem is that one is not limited to examining only conditional expectations of Y: one can re-weight the data and examine other aspects of the distribution, such as quantiles, variances, etc. by computing the desired statistic with the appropriate weight. See (DiNardo et al., 1996; DiNardo and Lemieux, 1997; Biewen, 1999; Firpo, 2007) for discussion and applications.

5 Program Evaluation: Lessons and Challenges

This chapter provides a systematic assessment of a selection of commonly employed program evaluation approaches. We adopt a perspective that allows us to consider how the Regression Discontinuity Design – an approach that has seen a marked increase in use over the past decade – relates to other well-known research designs. In our discussion, we find it helpful to make two distinctions. One is between the descriptive goals of an ex post evaluation, and the predictive goals of an ex ante evaluation. And the other is between two kinds of statistical conditions needed to make causal inference – 1) descriptions of our institutional knowledge of the program assignment process, and 2) structural assumptions – some that have testable restrictions, and others that will not – that do not come from our institutional knowledge, but rather stem from conjectures and theories about individual behavior; such structural assumptions necessarily restrict the set of models of

behavior within which we can consider the causal inference to be valid.

In our discussion, we provide three concrete illustrations of how the goals of ex post and ex ante evaluations are quite complementary. In the case of the randomized experiment with perfect compliance, highly credible estimates can be obtained for program effects for those who selected to be a participant in the study. Through the imposition of a number of structural assumptions about the nature of the economy, one can draw a precise link between the experimentally obtained treatment effect and a particular policy parameter of interest – the aggregate impact of a wide-spread "scaling" up of the program. In the case of the randomized experiment with imperfect compliance, one can make highly credible inferences about program effects, even if the obtained treatment effect is a weighted average. But with an additional functional form assumption (that is by no means unusual in the applied literature), one can extrapolate from a Local Average Treatment Effect to the Average Treatment Effect, which might be the "parameter of interest" in an ex ante evaluation. Finally, in the case of the RD design, one can obtain highly credible estimates of a weighted average treatment effect, which in turn can be viewed as an ingredient to an extrapolation for the Treatment on the Treated parameter.

Our other observation is that "D"-conditions and "S"-conditions are also quite complementary. On the one hand, for the designs we examine above "D-"conditions are generally necessary (even if not sufficient) to isolate the component of variation in program status that is "as good as randomly assigned". When they are not sufficient, "S"-conditions are needed to fill in the missing pieces of the assignment process. Furthermore, in our three illustrations, only with "S-" conditions can any progress be made to learn about other parameters of interest defined by an ex ante evaluation problem. Thus, our three examples are not meant to be definitive, but rather illustrative of how research designs dominated by D-conditions could supply the core ingredients to ex ante evaluations that are defined by S-conditions. In our view, this combination seems promising.

More importantly, what is the alternative? There is no a priori reason to expect that the variation that we may be able to isolate as "effectively randomized", to be precisely the variation required to identify a particular policy proposal of interest, particularly since what is "of interest" is subjective, and researcher-dependent.[47] That is, in virtually any context – experimental or non-experimental – the effects we *can* obtain will not exactly match what we *want*. The alternative to being precise about the sub-population for whom the effects are identified is to be imprecise about it. And for conducting an ex ante evaluation,

[47] Heckman and Vytlacil (2005) make the point that if the propensity score has limited support (e.g. including discrete support), marginal treatment effects cannot be identified in certain areas, and certain policy parameters of interest are also not identified.

the alternative to using an extrapolation where the leading term is a highly credible experimental/quasi-experimental estimate is to abandon that estimate in favor of an extrapolation in which the leading term is an estimate with questionable or doubtful internal validity. Similarly, even if the assumptions needed for extrapolation involve structural assumptions that require an uncomfortable suspension of disbelief, the alternative to being precise in specifying those assumptions, is to be imprecise about it and make unjustified generalizations, or to abandon the ex ante evaluation question entirely.

We conclude with some speculation on what could be fruitful directions for further developing strategies for the ex post evaluation problem. One observation from our discussion is that both the Sharp RD and Fuzzy RD are representations of the general self-selection problem, where agents can take actions to influence their eligibility or participation in a program. What allows identification – and indeed the potential to generate randomization from a non-experimental setting – is our *knowledge* of the threshold, and the *observability* of the "latent variable" that determines the selection. Turning that on its head, we could view *all* selection problems with a latent index structure as inherently Regression Discontinuity designs, but ones for which we do *not* perfectly observe the latent selection variable (or the cutoff). But what if we have partial institutional knowledge on the assignment process? That is, even if we don't measure V (from Sections 3.3 and 3.4), what if we observe a reasonable proxy for V? Can that information be used?

On a related point, our presentation of various research designs has a "knife-edge" quality. The designs in Section 3 are such that P^* or p_{z1} have point-mass distributions, or we required *every* individual to have a continuous density for V. When those conditions held, we argued that the effects would be highly credible, with strong, testable implications. But in Section 2.2.1, we argued that when we do *not* have specific knowledge about the assignment process, the designs will tend to yield more questionable inferences, because there will be an increase in plausible alternative specifications often if with very little to guide us as to the "preferred" specification. Does a middle-ground exist? Are there situations where our knowledge of the assignment process tells us that P^* or p_{z1}, while not distributed as a mass-point, has *small* variance? Might there be ways to adjust for these "minor" departures from "as good as randomized"?

Finally, another lesson from the RD design is how much is gained from actually knowing something about the treatment assignment process. It is intuitive that when one actually knows the rule that partially determines program status, and one observes the selection rule variable, that should help matters. And it is intuitive that if program assignment is a complete "black box" – as is often the case when researchers invoke a "selection on observables"/matching approach – we will be much less confident about those program

effects; one ought to be a bit skeptical about strong claims to the contrary. Since most programs are at least partially governed by some eligibility rules, the question is whether there are some other aspects of those rules – that go beyond discontinuities or actual random assignment – from which we can tease out credible inferences on the programs' causal impacts.

References

Abadie, Alberto, Joshua D. Angrist, and Guido Imbens, "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, January 2002, *70* (1), 91–117.

Abbring, Jaap H. and James J. Heckman, "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 6 of *Handbook of Econometrics*, Elsevier, December 2007, chapter 72.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 1990, *80* (3), 313–336.

_ , "Treatment effect heterogeneity in theory and practice," *Economic Journal*, 03 2004, *114* (494), C52–C83.

_ **and Alan B. Krueger**, "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3-A of *Handbooks in Economics*, New York: Elsevier Science, 1999, chapter 23, pp. 1277–1366.

_ **and Jörn-Steffen Pischke**, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," NBER Working Papers 15794, National Bureau of Economic Research, Inc March 2010.

_ **and Victor Lavy**, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, May 1999, *114* (2), 533–575.

_ **and William N Evans**, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, June 1998, *88* (3), 450–77.

_ **, Guido W. Imbens, and Donald B. Rubin**, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Angrist, Joshua, Eric Bettinger, and Michael Kremer, "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, June 2006, *96* (3), 847–862.

Ashenfelter, Orley, "Estimating the effect of training programs on earnings.," *Review of Economics and Statistics*, 1978, *60* (1), 47–57.

_ **and David Card**, "Time Series Representations of Economic Variables and Alternative Models of the Labour Market," *Review of Economic Studies*, 1982, *49* (5), 761–81.

_ **and** _ , "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 1985, *67*.

_ **and Mark W. Plant**, "Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs," *Journal of Labor Economics*, 1990, *8* (1), S396–S415.

Barnow, B., G. Cain, and A. Goldberger, "Issues in the Analysis of Selectivity Bias," *Evaluation Studies Review Annual*, 1976, *5*, 43–59.

Biewen, Martin, "Measuring the Effects of Socio–Economic Variables on the Income Distribution: An Application to the East German Transition Process," Discussion Paper Series, Ruprecht–Karls–Universität, Heidelberg, Germany March 1999.

Black, Sandra, "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, May 1999, *114* (2), 577–599.

Busso, Matias, John DiNardo, and Justin McCrary, "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," Unpublished Working Paper, University of Michigan, Ann Arbor, MI September 19 2008.

_ , _ , and _ , "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," Working Paper 3998, Institute for the Study of Labor (IZA) February 2009.

Campbell, Donald T. and Thomas D. Cook, *Quasi–Experimentation: Design and Analysis for Field Settings*, first ed., Chicago: Rand McNally College Publishing Company, 1979.

Card, David and Alan B. Krueger, *Myth and measurement : the new economics of the minimum wage*, Princeton, N.J.: Princeton University Press, 1995.

_ , Carlos Dobkin, and Nicole Maestas, "Does Medicare Save Lives?," *Quarterly Journal of Economics*, 2009, *124(2)*, 597–636.

_ , _ , and _ , "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," *American Economic Review*, forthcoming 2009.

Cook, T.D., ""Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics," *Journal of Econometrics*, February 2008, *142* (2), 636–654.

Cox, D. R., *Planning of Experiments*, New York: Wiley, 1958.

Deaton, Angus S., "Instruments of Development: randomization in the tropics and the search for the elusive keys to development," *Proceedings of the British Academy*, October 7 2008, *162*, 123–160. Keynes Lecture, British Academy.

DiNardo, John and David S. Lee, "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001," *Quarterly Journal of Economics*, November 2004, *119* (4), 1383 – 1441.

_ and Thomas Lemieux, "Diverging Male Wage Inequality in the United States and Canada, 1981–1988: Do Institutions Explain the Difference?," *Industrial and Labor Relations Review*, August 1997.

_ , Nicole Fortin, and Thomas Lemieux, "Labor Market Institutions and The Distribution of Wages, 1973-1993: A Semi-Parametric Approach," *Econometrica*, September 1996, *64* (5), 1001–1045.

Fan, Jianqing and Irene Gijbels, *Local Polynomial Modelling and Its Applications*, New York: Chapman and Hall, 1996.

Fang, Hanming, Michael Keane, Ahmed Khwaja, Martin Salm, and Daniel Silverman, "Testing the Mechanisms of Structural Models: The Case of the Mickey Mantle Effect," *American Economic Review*, May 2007, *97* (2), 53–59.

Fernández-Villaverde, Jesús, "The Econometrics of DSGE Models," Working Paper 14677, National Bureau of Economic Research January 2009.

Field, Erica, "Entitled to Work: Urban Property Rights and Labor Supply in Peru," *The Quarterly Journal of Economics*, November 2007, *122* (4), 1561–1602.

Firpo, Sergio, "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, January 2007, *75* (1), 259 – 276.

_ **, Nicole Forin, and Thomas Lemieux**, "Decomposition Methods in Economics," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 4, Amsterdam: North Holland, 2010.

Fisher, Sir Ronald Aylmer, *Design of Experiments*, Edinburgh, London: Oliver and Boyd, 1935.

_ , *Design of Experiments*, 8th ed., Edinburgh, London: Oliver and Boyd, 1966. First edition published in 1935.

Freedman, David A., "A Note on Screening Regression Equations," *The American Statistician*, 1983, *37* (2), 152–155.

Guttman, Robert, "Job Training Partnership Act: new help for the unemployed," *Monthly Labor Review*, March 1983, pp. 3–10.

Haavelmo, Trygve, "The Probability Approach in Econometrics," *Econometrica*, 1944, *12*, iii–115.

Hacking, Ian, *The Logic of Statistical Inference*, Cambridge: Cambridge University Press, 1965.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estim ation of Average Treatment Effects," *Econometrica*, March 1998, *66* (2), 315–331.

_ **, Petra Todd, and Wilbert Van der Klaauw**, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, January 2001, *69* (1), 201–209.

Hearst, Norman, Tom B. Newman, and Stephen B. Hulley, "Delayed Effects of the Military Draft on Mortality:A Randomized Natural Experiment," *New England Journal of Medicine*, March 6 1986, *314*, 620–624.

Heckman, James J., "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 1974, *42* (4), 679–694.

_ , "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 1976, *5* (4), 475–492.

_ , "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 1978, *46*, 931–960.

_ , "Randomization and Social Policy Evaluation," Working Paper 107, National Bureau of Economic Research July 1991.

_ , "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics*, February 2000, *115* (1), 45–97.

_ , "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 2001, *109* (4), 673–748.

_ **and Bo E. Honore**, "The Empirical Content of the Roy Model," *Econometrica*, 1990, *58* (5), 1121–1149.

_ **and Edward J. Vytlacil**, "Local Instrumental Variables," in Cheng Hsiao, Kimio Morimune, and James L. Powell, eds., *Nonlinear statistical modeling : proceedings of the thirteenth International Symposium in Economic Theory and Econometrics : essays in honor of Takeshi Amemiya*, number 13. In 'International Symposia in Economic Theory and Econometrics.', Cambridge University Press, 2001, chapter 1.

_ **and** _ , "Policy-Relevant Treatment Effects," *The American Economic Review*, May 2001, *91* (2), 107–111. Papers and Proceedings of the Hundred Thirteenth Annual Meeting of the American Economic Association.

_ **and** _ , "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, May 2005, *73* (3), 669–738.

_ **and** _ , "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, 1 ed., Vol. 6B, Elsevier, 2007, chapter 70.

_ **and** _ , "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 6 of *Handbook of Econometrics*, Elsevier, December 2007, chapter 71.

_ **and Jeffrey A. Smith**, "Evaluating the Welfare State," NBER Working Papers 6542, National Bureau of Economic Research, Inc May 1998.

_ **and Richard Robb Jr.**, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press, 1985.

_ **and Sergio Urzua**, "Comparing IV With Structural Models: What Simple IV Can and Cannot Identify," Working Paper 14706, National Bureau of Economic Research February 2009.

_ **, H. Ichimura, and Petra Todd**, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, April 1998, *65* (2), 261–294.

_ **, Justin L. Tobias, and Edward J. Vytlacil**, "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, October 2001, *68* (2), 210–223.

_ **,** _ **, and** _ , "Simple Estimators for Treatment Parameters in a Latent Variable Framework," *Review of Economics and Statistics*, August 2003, *85* (3), 748–755.

_ **, Robert J. LaLonde, and James A. Smith**, "The Economics and Econometrics of Active Labour Market Programmes," in "The Handbook of Labour Economics," Vol. III North–Holland Amsterdam 1998.

_ **, Sergio Urzua, and Edward J. Vytlacil**, "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, August 2006, *88* (3), 389–432.

Hirano, Keisuke, Guido Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, December 1986, *81* (396), 945–960.

Imbens, Guido and Joshua Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, March 1994, *62* (2), 467–476.

_ **and Thomas Lemieux**, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, February 2008, *142* (2), 615–635.

_ **and** _ , "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, *Forthcoming* 2008.

Imbens, Guido W., "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," Working Paper 14896, National Bureau of Economic Research April 2009.

Keane, Michael P., "Structural vs. atheoretic approaches to econometrics," *Journal of Econometrics*, 2009, *In Press, Corrected Proof*, –.

_ **and Kenneth I. Wolpin**, "Exploring The Usefulness Of A Nonrandom Holdout Sample For Model Validation: Welfare Effects On Female Behavior," *International Economic Review*, November 2007, *48* (4), 1351–1378.

Lee, David S., "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, February 2008, *142* (2), 675–697.

_ **and David Card**, "Regression discontinuity inference with specification error," *Journal of Econometrics*, 2008, *142* (2), 655 – 674. The regression discontinuity design: Theory and applications.

_ **and Thomas Lemieux**, "Regression Discontinuity Designs in Economics," Working Paper 14723, National Bureau of Economic Research February 2009.

Lehmann, Erich Leo, *Testing Statistical Hypotheses*, New York: John Wiley & Sons, Inc., 1959.

_ **and Joseph Lawson Hodges Jr.**, *Basic Concepts of Probability and Statistics*, San Francisco: Holden-Day, 1964.

Lemieux, Thomas and Kevin Milligan, "Incentive Effects of Social Assistance: A Regression Discontinuity Approach," *Journal of Econometrics*, February 2008, *142* (2), 807–828.

Lucas, Robert Jr, "Econometric policy evaluation: A critique," *Carnegie-Rochester Conference Series on Public Policy*, January 1976, *1* (1), 19–46.

Maddala, G. S., *Limited-dependant and qualitative variables in Econometrics*, Cambridge University Press, 1983.

Manning, Alan, *Monopsony in Motion: Imperfect Competition in Labor Markets*, Princeton, NJ: Princeton University Press, April 2003.

Marschak, Jacob, "Economic Measurements for Policy and Prediction," in William C. Hood and Tjalling C. Koopmans, eds., *Studies in Econometric Method*, John Wiley and Sons New York 1953, pp. 1–26.

Mayo, Deborah G., *Error and the Growth of Experimental Knowledge* Science and Its Conceptual Foundations, Chicago: University of Chicago Press, 1996.

McCall, Brian Patrick and John Joseph McCall, *The economics of search*, Routledge, London ; New York :, 2008.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, February 2008, *142* (2), 698–714.

_ **and Heather Royer**, "The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Laws Using Exact Date of Birth," Unpublished Working Paper, University of California Berkeley *Forthcoming* 2010.

McFadden, Daniel, Antti Talvitie, and Associates, "Demand Model Estimation and Validation," Urban Travel Demand Forecasting Project UCB-ITS-SR-77-9, The Institute of Transportation Studies, University of California, Irvine and University of California, Berkeley 1977. Phase 1 Final Report Series, Volume V.

Millimet, Daniel L. and Rusty Tchernis, "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies," *Journal of Business and Economic Statistics*, 2009, *27* (3), 397–415.

Oreopoulos, Phillip, "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, March 2006, *96* (1), 152–175.

Orr, Larry, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling, "Moving to Opportunity Interim Impacts Evaluation," *Final Report, U.S. Department of Housing and Urban Development*, 2003.

Pagan, A. and A. Ullah, *Nonparametric Econometrics*, Cambridge University Press, New York, 1999.

Powell, James L., "Estimation of Semiparametric," in Robert Engle and Daniel McFadden, eds., *Handbook of Econometrics*, Vol. 4, Amsterdam: North Holland, 1994.

Quandt, Richard E., "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, 1958, *53* (284), 873–880.

_ , "A New Approach to Estimating Switching Regressions," *Journal of the American Statistical Association*, 1972, *67* (338), 306–310.

Reiss, Peter C. and Frank A. Wolak, "Structural Econometric Modeling: Rationales and Examples from Industrial Organization," in James J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, Vol. 6 of *Handbook of Econometrics* Elsevier 2007.

Robins, Philip K., "A Comparison of the Labor Supply Findings from the Four Negative Income Tax Experiments," *The Journal of Human Resources*, 1985, *20* (4), 567–582.

Rosen, Sherwin, "The theory of equalizing differences," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. 1 of *Handbook of Labor Economics*, Elsevier, December 1987, chapter 12, pp. 641–692.

Rosenbaum, Paul and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 1983, *70* (1), 41–55.

Rosenzweig, Mark R. and Kenneth I. Wolpin, "Natural 'Natural Experiments' In Economics," *Journal of Economic Literature*, December 2000, *38* (4), 827–874.

Roy, A., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 1951, *3* (2), 135–146.

Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, October 1974, *66* (5), 688–701.

_ , "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, December 1986, *81* (396), 961–962.

_ , "[On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 1990, *5* (4), 472–480.

Rubin., Donald B. and N. Thomas, "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 1996, *52*, 249.

Rust, John, "Comments on: by Michael Keane," *Journal of Econometrics*, 2009, *In Press, Corrected Proof*, –.

Splawa-Neyman, Jerzy, D. M. Dabrowska, and T. P. Speed, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1990, *5* (4), 465–472.

_ , K. Iwaszkiewicz, and St. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation," *Supplement to the Journal of the Royal Statistical Society*, 1935, *2* (2), 107–180.

Taber, Christopher and Eric French, "Identification of Models of the Labor Market," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 4, Elsevier Science, 2010.

Tesfatsion, Leight, "Introductory Notes on Complex Adaptive Systems and Agent-Based Computational Economics," Technical Report, Department of Economics, Iowa State University January 2007. `http://www.econ.iastate.edu/classes/econ308/tesfatsion/bat1a.htm`.

Thistlethwaite, Donald L. and Donald T. Campbell, "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment," *Journal of Educational Psychology*, December 1960, *51*, 309–317.

Todd, Petra and Kenneth Wolpin, "Ex Ante Evaluation of Social Programs," PIER Working Paper Archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania 2006.

Van der Klaauw, Wilbert, "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, November 2002, *43* (4), 1249–1287.

_ , "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, June 2008, *22* (2), 219–245.

Van der Klaauw, Wilbert, "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *LABOUR*, 06 2008, *22* (2), 219–245.

Windrum, Paul, Giorgio Fagiolo, and Alessio Moneta, "Empirical Validation of Agent-Based Models: Alternatives and Prospects," *Journal of Artificial Societies and Social Simulation*, 2007, *10.*

Wolpin, Kenneth I., "Ex Ante Policy Evaluation, Structural Estimation and Model Selection," *American Economic Review*, May 2007, *97* (2), 48–52.

Table 1. Extrapolating from LATE to ATE

$$ATE = LATE - (\rho_1\sigma_1 - \rho_0\sigma_0) \cdot \underbrace{\frac{\phi\left(\Phi^{-1}\left(1 - E\left[D|Z=1\right]\right)\right) - \phi\left(\Phi^{-1}\left(1 - E\left[D|Z=0\right]\right)\right)}{E\left[D|Z=1\right] - E\left[D|Z=0\right]}}_{*}$$

	Abadie et al. 2002	Angrist et al. 2002	Angrist and Evans 1998	Field 2007
Context	Effect of training on trainee earnings	Vouchers for private schooling in Colombia	Effect of childbearing on labor supply	Titling program for urban squatters in Peru
Outcome (Y)	Earnings	Schooling	Worked for pay	Total weekly work hours
Treatment (D)	Training	Scholarship use	More than two children	Title possession and report of experiencing change in tenure security
Instrument (Z)	Offer of training	Received a voucher	First two children are of the same sex	Program in Neighborhood
(1) **Mean of Y**	19147	7.400	0.565	103.69
	[19540]	[1.099]	[0.496]	[77.68]
(2) **E[Y|D=1]-E[Y|D=0]**	3970	0.292	-0.121	-1.07
(3) **LATE** $= \frac{E[Y\|Z=1]-E[Y\|Z=0]}{E[D\|Z=1]-E[D\|Z=0]}$	1825	0.168	-0.132	31.59
(4) $\rho_1\sigma_1$	-3646	0.00012	0.0265	-27.01
(5) $\rho_0\sigma_0$	2287	0.326	-0.0088	-9.96
(6) $\rho_1\sigma_1 - \rho_0\sigma_0$ **:** (4)-(5)	-5932	-0.326	0.035	-17.05
(7) *** Term**	0.57	-0.221	0.249	-0.30
(8) **Selection Corretion Term:** -(6)x(7)	3400	-0.072	-0.0088	-5.07
	(882)***	(0.032)**	(0.0082)	(16.55)
(9) **ATE:** (3)+(8)	5225	0.096	-0.141	26.52

Note: Standard deviations are in brackets. Standard errors (in parentheses) are calculated using the delta method.

Figure 1. Density of Assignment Variable Conditional on $W = w$, $U = u$