Viewpoint: Estimating the causal effects of policies and programs

Jeffrey Smith *University of Michigan, NBER, IZA and CESifo*
Arthur Sweetman *McMaster University and IZA*

Abstract. Estimation, inference and interpretation of the causal effects of programs and policies have all advanced dramatically over the past 25 years. We highlight three particularly important intellectual trends: an improved appreciation of the substantive importance of heterogeneous responses and of their methodological implications, a stronger focus on internal validity brought about by the "credibility revolution," and the scientific value that follows from grounding estimation and interpretation in economic theory. We discuss a menu of commonly employed partial equilibrium approaches to the identification of causal effects, emphasizing that the researcher's central intellectual contribution always consists of making an explicit case for a specific causal interpretation given the relevant economic theory, the data, the institutional context and the economic question of interest. We also touch on the importance of general equilibrium effects and full cost–benefit analyses.

Résumé. *Point de vue: Sur l'estimation des effets causatifs des politiques et programmes.* Dans le monde de l'estimation, l'inférence et l'interprétation des effets causatifs des programmes et des politiques, il y a eu des progrès dramatiques au cours des derniers 25 ans. Les auteurs soulignent trois tendances intellectuelles particulièrement importantes : une appréciation améliorée de l'importance substantielle des réponses hétérogènes et de leur importance méthodologique, une focalisation plus robuste sur la validité interne engendrée par la « révolution de la crédibilité », et la valeur scientifique qui découle d'un ancrage de l'estimation et de l'interprétation dans la théorie économique. On discute un éventail d'approches d'équilibre partiel à l'identification des effets causatifs, mettant au premier plan que la contribution intellectuelle centrale du chercheur consiste à bâtir un argumentaire explicite pour une interprétation causale spécifique compte tenu de la théorie économique pertinente, des données, du contexte institutionnel, et de la question économique d'intérêt. On mentionne aussi l'importance des effets d'équilibre général et des analyses de tous les coûts et avantages.

JEL classification: C18, C21, C26, C50, C90

We thank David Green, Lonnie Magee, Craig Riddell and Mike Veall for helpful comments and the editors for their patience with us. This article draws on a piece presented in August 2009 at the roundtable on Strengthening Evidence-Based Policy in the Australian Federation, hosted in Canberra by the Australian Productivity Commission. That paper is available at pc.gov.au/research/supporting/strengthening-evidence. Any and all errors of omission or commission, alas, remain our responsibility.
Corresponding author: Arthur Sweetman, arthur.sweetman@mcmaster.ca

Canadian Journal of Economics / Revue canadienne d'économique, Vol. 49, No. 3
August 2016. Printed in Canada / Août 2016. Imprimé au Canada

0008–4085 / 16 / 871–905 /

1. Introduction

Methods for estimating, making inferences about and interpreting the causal effects of programs and policies have advanced dramatically over the past 25 years. This paper briefly reviews and comments on these developments, with an emphasis on issues we think have received insufficient attention in other discussions.[1]

Three major intellectual trends within applied economics frame our discussion. The first is serious consideration of heterogeneity in responses to programs and policies, which we consider in section 2. Heterogeneous responses arise for many reasons, including the common and prosaic practice of reducing heterogeneous programs and policies to binary indicator variables. In the "essential heterogeneity" world of Heckman (2001), agents and gatekeepers make policy and program participation choices in light of (perhaps quite limited) knowledge of heterogeneous responses. Throughout the paper we consider the implications of heterogeneous responses for the identification, estimation and interpretation of estimated causal effects.

The second trend is the "credibility revolution," which seeks to increase the internal validity of estimates of program and policy effects via reliance on carefully considered identification strategies. This revolution encompasses both the very rapid increase in the use of social and field experiments as well as the increase in attention to the sources of exogenous variation employed in techniques such as instrumental variables. More broadly, it embodies a greater emphasis on internal validity relative to other concerns. Works such as Angrist and Pischke (2009), Rubin (2008) and Imbens (2010) embody the credibility revolution worldview.

The third trend is the intellectual revival of "structural" methods. Loosely speaking, applying the structural approach means writing down an explicit economic model and estimating (or sometimes calibrating) the parameters of the model. The formal model facilitates interpretation and, to the extent that it achieves the goal of plausible policy invariance, provides a theoretically grounded approach to consideration of policies not present in the data. This revival has many sources: in part it represents a reaction against the excesses of the credibility revolution.[2] It also reflects improvements in computing technology that allow the estimation of more sophisticated structural models as well as conceptual developments related to estimation, interpretation and the incorporation of heterogeneity. The new wave of structural research also responds positively to the credibility revolution via closer attention to sources of identification, be they model-based (i.e., largely dependent upon economic theory and/or functional

1 We oversample examples from the labour economics, health economics and economics of education literatures, as we know these literatures best. However, the underlying issues know no subfield boundaries. We do not delve deeply into technical econometrics; instead, we point readers to the many excellent books and reviews that cover that material, such as Abbring and Heckman (2007), Heckman and Vytlacil (2007a, b), Imbens and Wooldridge (2009) and Imbens and Rubin (2015). Finally, we do not address Bayesian approaches.

2 One of us fondly recalls initiating the following exchange at a conference: Question: "What model could explain your results?" Answer: "We think of our paper as a theory-free zone."

form assumptions) and/or design-based (i.e., largely based on a—potentially theoretically motivated—research design involving exogenous variation). Heckman (2001, 2005, 2010) makes the intellectual case for the modern structural view.[3]

We view heterogeneous treatment effects as obvious and long overdue, and we view strong and explicitly defended identification as a complement to, rather than a substitute for, interpretation based on coherent economic frameworks. This does not mean that every paper must have a structural, or even a formal, model. Full-blown structural models require specific expertise and consume a great deal of research time and should be subjected to ex ante intellectual cost–benefit tests. Moreover, economists have certain theories about the value of the division of labour. We therefore see the approaches as complements at the disciplinary level and so resist calls for every empirical economist to become a part-time producer of formal theory. Further, we strongly reject the view that economists should not bother trying to answer questions for which the extant data and institutions provide no strong design-based identification. Instead, we argue that economists should try to operate at the frontier in a space defined by the credibility of identification and the substantive importance of the topic and should recognize that quality of evidence takes a continuous rather than a binary form.[4]

Section 3 lays out our viewpoint (as promised in the title!) regarding the implications of these three complementary trends for the practice of program and policy evaluation by economists. To do so, it marches through the standard toolkit of partial equilibrium applied econometric identification strategies—randomization, conditional independence (a.k.a. "selection on observed variables"), instrumental variables, bias stability (a.k.a. "natural experiments") and regression discontinuity—describing the substantive implications of each class of assumptions for estimation and interpretation and discussing how to make an explicit case for a specific causal interpretation given the economic theory, the data, the institutional context and the evaluation question of interest. We view the making of such substantive cases as the primary intellectual task of the empirical researcher seeking to make causal claims.

Section 4 considers two related alternative views of recent developments that focus solely on design-based approaches: hierarchies of evidence and "magic bullet" theories regarding particular estimators. These views attempt to take an epistemological shortcut around the hard work of making a compelling case relevant to a particular evaluation context. We argue for rejecting these views. Sections 5 and 6 consider equilibrium effects and cost–benefit analyses, respectively. These two topics nicely illustrate our broader point regarding the complementary nature of credible design-based identification and economic theory. Section 6 concludes.

3 The structural crowd has its excesses as well, such as zealous arguments for the proposition that a probit model is in some meaningful sense more structural than a linear probability model. We also view the "Type I Extreme Value is the New Normal" T-shirt through the lens of Goldberger (1983).

4 Although we focus on causality here, we recognize the value of empirical research that focuses on (sometimes sophisticated conditional) covariances, trends over time and the like; "descriptive" (that is, explicitly non-causal) should not be a derisive term in empirical research.

2. Parameters of interest

The standard potential outcomes framework provides a valuable way in which to think about causal parameters of interest. To begin, define two outcomes and a treatment indicator:

Y_{1i} denotes the outcome for unit "i" with treatment
Y_{0i} denotes the outcome for unit "i" without treatment
$T_i \in \{0, 1\}$ indicates treatment status for unit "i"

We limit ourselves to binary treatments for simplicity. The general case of discrete multi-valued treatments follows easily (sometimes); the still more general case of continuous treatments can add non-trivial complications. Because a given unit can experience only one of the two outcomes, we have the observation equation:

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i},$$

where Y_i denotes the observed outcome for unit "i". Following the literature, and to emphasize the generality of the framework, we use "units" as a generic term for participants and "non-participants" to emphasize that programs may serve, say, individuals, firms or local governments. We use "treatment" as a generic term for programs and policies. Two other bits of notation will prove useful later on: let X_i denote a vector of exogenous conditioning variables (but see the discussion at footnote 10) and let Z_i denote a variable that induces exogenous variation in treatment status.

For treated units, we observe the treated outcome while the untreated outcome remains counterfactual. For the untreated units, we observe the untreated outcome while the treated outcome remains counterfactual. The difference between the treated outcome and the untreated outcome defines the always-unobserved treatment (or "causal") effect for each unit:

$$\delta_i = Y_{1i} - Y_{0i}$$

Up to this point, we have implicitly interpreted the potential outcome of one unit as unaffected by the treatment status of other units. The literature calls this the Stable Unit Treatment Value Assumption (SUTVA), which is closely allied to what economists term partial equilibrium. We relax this assumption, which rules out general equilibrium effects, in section 5.

The literature focuses on particular averages of δ_i, where the choice of which average depends on the academic or policy question of interest, subject to constraints following from the identification strategy and the data that we discuss in section 3. The most common causal estimand is the Average Treatment Effect on the Treated (ATET, or sometimes TT or TOT), given by $\text{ATET} = E(Y_1 - Y_0 | T = 1)$; we suppress the "$i$" subscript inside the expectations operator both here and elsewhere. This parameter informs a cost–benefit analysis that addresses the question of whether to keep or scrap a program in its present form. Another

common estimand is the Average Treatment Effect (ATE), defined as ATE = $E(Y_1 - Y_0)$. The ATE equals the expected impact in the entire population of eligible units, whether or not they actually participate. This parameter informs a cost–benefit analysis that considers a mandatory program.

For voluntary programs, impacts at the margin of participation inform cost–benefit analyses regarding marginal expansions or contractions in the number of units treated. The definition of marginal will depend on the program context: it might represent volunteers close to indifferent between participating and not, or it may represent participants on the margin of selection by a caseworker or other gatekeeper. Particular sources of exogenous variation in participation, Z, each define their own set of marginal units that are associated with parameters such as the Local Average Treatment Effect (LATE) and the Marginal Treatment Effect (MTE), as we discuss in section 3.3.

The marginal distributions of outcomes $F(Y_{1i})$ and $F(Y_{0i})$ suffice to identify all of the treatment effect parameters considered so far. Another class of parameters of interest requires the joint distribution $F(Y_{1i}, Y_{0i})$. These include the impact variance, $\text{var}(\delta_i) = \text{var}(Y_{1i} - Y_{0i})$, and the fraction of units with a positive impact, $\Pr(\delta_i > 0)$. Identification of these parameters requires assumptions even in the presence of experimental data, because experimental data provide only the two marginal distributions, not the link between them implicit in the joint distribution. Because this class of parameters has received (too) little attention in applied work, we do not discuss it further but instead refer the interested reader to Djebbari and Smith (2008).

2.1. A simple Roy-inspired model

Consider a very simple model inspired by Roy (1951) and by the models in Heckman et al. (1999) and Barnow and Smith (2016). It adds to the potential outcomes framework a direct cost of treatment, call it C_i, which may include money costs, time costs and psychic costs of treatment that may be positive or negative. Assume, for the moment, that units maximize by choosing to participate when $Y_{1i} - C_i > Y_{0i}$.

This simple model embodies three different types of non-random selection. First, holding $Y_{1i}$and C_i fixed, units with high values of Y_{0i} do not participate and units with low values of Y_{0i} do participate; in many contexts, this represents units with a high opportunity cost declining to participate. Back in the dinosaur days of the common effect model, in which $Y_{1i} - Y_{0i} = \delta$, a constant for all "i", selection on the untreated outcome represented the primary source of applied econometric concern as in the classical bivariate normal selection model studied by Heckman (1979). Second, holding C_i fixed, we observe selection on the treatment effect $Y_{1i} - Y_{0i}$ whereby units with high impacts select into treatment. Björklund and Moffitt (1987) add selection based on the magnitude of the individual-specific impact to the bivariate normal model.[5] Third, holding the potential outcomes fixed, we expect selection into the program based on C_i as units with low costs select into

5 Blundell et al. (2005) provide an especially clear explication and application.

the program. As discussed in section 3.3, many instrumental variables analyses rely on cost-based instruments.

As much as anything, this simple Roy-inspired model points to the limitations of much research that focuses on estimating causal impacts. Standard economic models assume that choices depend not on potential outcomes but on the expected utility associated with those outcomes; that expected utility typically depends on both the outcomes studied and on other, unmeasured, outcomes. Similarly, the costs perceived by units pondering treatment will likely include components, such as psychic costs, not available to the evaluator. Even ordering the magnitudes of the various causal estimands requires assumptions regarding the correlations between Y_{0i}, Y_{1i} and C_i. For example, consider the simple (but often empirically implausible) assumption of independence of costs from potential outcomes. In this case, we have ATET > ATE due to positive selection on impacts; in other words, the average treatment effect on the treated exceeds the average treatment effect in the population. Similarly, ATNT < ATE, where ATNT equals the average treatment effect among the non-treated, due to selection out of treatment by those with relatively low treatment effects. Finally, ATE < MATE < ATET, where MATE, the Marginal Average Treatment Effect, corresponds to individuals just indifferent to participation, i.e., with $C_i \approx Y_{1i} - Y_{0i}$, whereas the ATET includes inframarginal participants whose treatment effect puts them far from the margin of indifference.

2.2. Parameters of interest compared to the naïve difference in treated and untreated group means

For the non-experimental case, consider estimation of the ATET, and then the ATE, in terms of: (i) the naïve expected mean difference between the observed treated outcome for the treated and the observed untreated outcome for the untreated, that is $E(Y_1|T=1) - E(Y_0|T=0)$, and (ii) various bias and/or selection terms. Start by writing:

$$\text{ATET} = E(Y_1 - Y_0|T=1) = E(Y_1|T=1) - E(Y_0|T=1).$$

Experiments produce $E(Y_0|T=1)$ by randomly forcing would-be $T=1$units to experience the untreated outcome, subject to various caveats that we discuss in section 3.1. To see the non-experimental case, add and subtract $E(Y_0|T=0)$ to yield:

$$\text{ATET} = [E(Y_1|T=1) - E(Y_0|T=0)] - [E(Y_0|T=1) - E(Y_0|T=0)].$$

The first term on the right-hand side is the population analogue of the naïve non-experimental mean difference estimator, while the second term consists of the expected difference in untreated outcomes between the participants and non-participants. This equation clearly illustrates the intuition from our Roy-inspired model that selection on the untreated outcome likely biases naïve comparisons of participant and non-participant outcomes. In particular, with costs independent of potential outcomes, the simple model predicts negative selection into

treatment based on Y_0, which implies a downward bias in the estimated treatment effect. A vast warehouse of evidence indicates that such selection matters in many contexts. The non-experimental identification strategies we discuss in sections 3.2 to 3.5 obtain estimates of the expected counterfactual outcome via econometric manipulation of the untreated units that removes such bias under specific assumptions.

Now, inspired by Heckman (1996a), contrast the naïve non-experimental mean difference to the ATE rather than the ATET. Extending the notation, let $E(V_1|T=j)=E(Y_1|T=j)-E(Y_1)$ denote the difference between the expected treated outcome among the treated ($j=1$) or the untreated ($j=0$) and the expected treated outcome in the population with $E(V_0|T=j)$ defined in the same way for the untreated outcome. We can interpret V as a residual of sorts. Then write the ATE in terms of the naïve non-experimental observed mean difference as:

$$\begin{aligned}\text{ATE} = E(Y_1)-E(Y_0) = [E(Y_1|T=1)-E(Y_0|T=0)] - [E(V_1|T=1) \\ - E(V_0|T=0)].\end{aligned}$$

Then, adding and subtracting $E(V_0|T=1)=E(Y_0|T=1)-E(Y_0)$ and manipulating yields:

$$\begin{aligned}\text{ATE} = E(Y_1)-E(Y_0) = [E(Y_1|T=1)-E(Y_0|T=0)] - E(V_1-V_0|T=1) \\ - [E(V_0|T=1)-E(V_0|T=0)].\end{aligned}$$

The ATE thus equals the naïve non-experimental mean difference plus two selection terms. The middle right-hand-side term captures selection into treatment based on the magnitude of the treatment effect. The rightmost term, in square brackets, captures selection into treatment based on the difference in the mean deviations between the population and the group-specific untreated outcomes. The Roy-inspired model with costs independent of impacts implies a positive value for the second term and a negative value for the third term, which means that the simple model does not sign the bias associated with using the naïve non-experimental mean difference to estimate the ATE in that context.

Treatment effects may vary with observed characteristics of the treated units, the program, or the broader environment. Djebbari and Smith (2008) call this "systematic" variation in treatment effects. For example, individuals who complete high school might benefit more from classroom-based occupational skills training courses than those who drop out, or an organizational intervention might work better in teaching hospitals than in other hospitals, or an active labour market program may have larger (or smaller) impacts during business cycle peaks than during troughs. Such variation interests researchers and policymakers for several reasons. Variation based on observed unit characteristics can guide efforts to use statistical treatment rules to target program services as in Manski (2004) and Lechner and Smith (2007). Variation in treatment effects allows researchers to test theories of treatment effect heterogeneity, as in Pitt et al. (2010). Finally,

and perhaps most importantly, treatment effect heterogeneity provides an input into discussions about the generalizability of findings across time, places and participant populations, what the literature calls their "external validity"; see, e.g., Hotz et al. (2005) and Muller (2015).

3. Partial equilibrium evaluation methods

This section lays out the standard econometric methods used to estimate the impact of interventions in a partial equilibrium context, i.e., in the context of the Stable Unit Treatment Value Assumption (SUTVA) introduced earlier. As standard statistical software packages now make it easy to use a variety of sophisticated estimators for partial equilibrium causal effects, the key intellectual exercises in most studies consist of: (1) justifying the SUTVA in the context at hand, (2) defending the chosen estimator's identification assumptions by making a case based on some combination of theory, contextual knowledge, specification tests and pre-existing empirical evidence and (3) correctly interpreting the results in light of the economics and context given the chosen identification strategy and estimator.

3.1. Social experiments

3.1.1. Random assignment and the selection bias problem

As noted above, experiments solve the selection bias problem by randomly forcing units that want to receive treatment to experience the untreated outcome instead, thereby ensuring that (in large samples) the treated and the control groups[6] have the same observed and unobserved characteristics.[7] In a randomized control trial, Z_i is the indicator for assignment to the treatment group; with perfect compliance, $T_i = Z_i$ while with partial compliance, Z_i imperfectly predicts T_i. In either case, Z_i is, by construction, uncorrelated with all observed and unobserved unit characteristics. This implies that $E(Y_0|Z=1) = E(Y_0|Z=0)$, which in turn implies that the difference in mean outcomes between the treatment and control groups estimates the ATET without bias when $Z_i = T_i$.

Probably Canada's best known large scale social experiments are the Self-Sufficiency Project (e.g., Ford et al. 2003, Lise et al. 2004, Card and Hyslop 2009 and Riddell and Riddell 2014), which randomly assigned single parents on income assistance in two provinces to a generous earnings subsidy program, and the

6 The economics literature uses the term "control group" rather loosely. We prefer a more precise categorization that restricts the term "control" to experimental or closely related situations where (active) control over the assignment of treatment status actually exists. In other situations we prefer the term "comparison group." It seems counterintuitive to apply the term control group to a situation where there is patently no control.

7 In some fields of economics, such as international trade, social experiments have not traditionally been undertaken and/or are not feasible, but this does not prevent experiments from serving as an intellectual touchstone. Also, while we focus in this section on random assignment addressing internal validity, it is random selection into the experimental group from the relevant population that addresses external validity.

"Mincome" guaranteed minimum income experiment in Manitoba considered in Hum and Simpson (1993) and Forget (2011). Less-expensive (though sometimes quite large scale) experiments have also occurred in the context of government operations (e.g., Warburton and Warburton 2002), and a large number of experiments can be found in health economics (e.g., Levin et al. 2007). There is tremendous potential to improve our understanding of the functioning of social, educational and other programs by introducing randomization or other sources of exogenous variation into program operations.

In the US, experiments have been applied to policy areas as diverse as health insurance, welfare-to-work programs, the handling of calls to the police reporting domestic violence, electricity pricing and abstinence-only sex education. Greenberg and Schroder (2004) document almost all of the earlier US experiments while Greenberg et al. (1999) present evidence on their characteristics. There has also been an explosion in experiments in developing countries; see Banerjee and Duflo (2009) on that score and see Levitt and List (2009) on the rise of experiments in economics more generally.

3.1.2. Issues with random assignment

Barnow (2010) says that random assignment is not a substitute for thinking. We agree that experiments present a more difficult evaluation challenge than their basic conceptual simplicity suggests. If well executed, experiments accomplish one very important thing: they eliminate the bias that results from non-random selection into treatment in partial equilibrium evaluations in a simple and compelling way. As noted by Heckman and Smith (2000), however, experiments remain subject to many of the other threats to internal validity that make empirical program evaluation using observational data so fraught with difficulties (and so interesting) such as outliers, survey non-response and attrition, error-filled and poorly documented administrative data, Hawthorne effects (when research subjects react to being observed) and John Henry effects (when experimental subjects react to being in the control group). In addition to these general issues, experiments also raise or exacerbate particular issues of internal and external validity.

Though not often discussed in economics, the simple fact of randomization does not in and of itself suffice for the identification of the ATET in an experiment with perfect compliance. A causal interpretation of the experimental estimand requires the additional assumption that randomized treatment assignment affects outcomes only via its effect on receipt of treatment. If we think of randomization as an ideal instrument in the sense of Heckman (1996a), this corresponds to the "exclusion restriction" assumption, which requires that an instrument affect outcomes only via its effect on the endogenous variable of interest. If, for example, explicit assignment to an experimental control group induces John Henry effects—reactive responses by control group members—then this assumption fails.

Outside economics, high quality pharmaceutical randomized control trials address this concern by being "blind" (the unit being randomized does not know and

thus cannot react to treatment status) or even better "double-blind" (neither the unit being randomized nor the program administrators know the treatment status). Failure of the exclusion restriction condition requires subtle interpretation; the medical literature reinterprets the control group as representing a placebo treatment rather than no treatment. Rather obviously, blinding presents insurmountable difficulties for most social experiments. For example, what would a placebo training program look like? Instead, researchers need to defend the plausibility of the exclusion restriction in their context and readers need to judge the credibility of these arguments.

In practice, experimental evaluations often fail to achieve full compliance. Not all treatment group members receive treatment or some control group members receive treatment ("cross-overs") or substitute something similar. In contexts with treatment group dropout and control group substitution the experiment ends up randomly assigning *access* to treatment, rather than treatment itself. Heckman et al. (2000) document the empirical importance of dropout and substitution in the context of US evaluations of employment and training programs. Researchers typically react to dropout and/or substitution in one of two ways. The first consists of reinterpreting the estimand as the mean impact of the offer of treatment (called "intention to treat" or ITT) rather than the mean impact of treatment itself. For policy purposes, this may have a greater interest than the ATET as researchers and governments can typically make offers of treatment but not compel it. The second consists of using random assignment as an instrument for receipt of treatment and interpreting the resulting estimand as a Local Average Treatment Effect (LATE), which corresponds to the average effect of treatment on those induced to change treatment status via randomization to the treatment group.[8]

Experimental evaluations may also have issues with external validity above and beyond those present in non-experimental evaluations. Random selection of eligible units for an experiment with a sufficiently large number of units ensures external validity. But in evaluations that require voluntary participation by local program sites, random assignment may raise the perceived cost of site participation and so lead to a non-random selection of sites, as described in Doolittle and Traeger (1990) for the US *Job Training Partnership Act* (JTPA) experiment. Similarly, the additional consent apparatus required for individual randomization may non-randomly deter participation, as in the UK Employment Retention and Advancement experiment studied in Sianesi (2014). On a different margin, experimental evaluations may require a program to dig deeper into its eligible population than it usually would in order to fill up the control group while still maintaining its normal scale of operations. Finally, randomized rather than deterministic access to treatment may deter complementary investments prior to

8 We say more about LATEs in our discussion of instrumental variables below. See Heckman, Smith et al. (1998) for more on dropout and section 5 of Heckman et al. (1999) for a more thorough general discussion. Kline and Walters (2015) provide an empirical example and pointers to the recent literature.

treatment or change the composition of participants by deterring the risk averse and attracting the risk loving.

A broader discussion embodied in the tryptic of Deaton (2010), Heckman (2010) and Imbens (2010) raises issues related to the nature and extent of the role of experimental methods within economics. In particular, it considers the importance of using methods that ensure compelling identification and so allow strong claims to internal validity. An exclusive focus on randomization (or on strong partial equilibrium designs more generally) may lead the discipline away from theoretical interpretability and external validity and may also lead to a focus on policy questions that experiments can easily address rather than a focus on the most substantively important policy questions.[9] As noted in the introduction, we think that experiments and structure often represent complements in the production of economic knowledge rather than substitutes. This comes through most strongly in the increasingly common analyses that explicitly combine the two, such as Lise et al. (2004) or Todd and Wolpin (2006). More broadly, as discussed in detail in Rothstein and von Wachter (2015), economic theory often provides an important guide to the questions worth studying using experiments and other strong designs, to the design of such studies, to the choice of outcome variables, to the choice of variables to measure and examine as potential mediators, to the conduct of cost–benefit analyses and so on. In sum, we think this debate ultimately makes the case for more thoughtful use of experimentation in the profession, not for its abandonment.

3.1.3. Variants of random assignment

Random assignment has many uses beyond the estimation of the ATET for use in cost–benefit analyses of whether to keep or drop programs. Additional uses address questions that sometimes possess equal or greater policy relevance and often avoid or reduce political, practical and ethical concerns related to a control group that receives little (if any) treatment. Consider some illustrative real world examples.

Black et al. (2003) document the clever use of randomization in the unemployment insurance (UI) system in Kentucky. Like other US states, Kentucky employs a statistical model to predict the fraction of their (traditionally) 26 weeks of UI benefit entitlement each new claimant will consume as a function of claimant and local area characteristics. Kentucky converts this continuous prediction into a discrete score from 1 to 20. In each local UI office in each week, the state assigns mandatory reemployment services to new UI claimants starting with those with the highest score and proceeding until it runs out of slots or claimants. In many

9 Or analysts may focus on strong design-based estimates even when they correspond to parameters of limited policy interest, as with the enthusiasm in the minimum wage literature for compelling estimates of the short-run labor demand response when the long-run demand response should guide most policy. As Sorkin (2015) shows, these two responses can differ quite substantially.

cases, for the marginal score (the one where the slots run out) the number of claimants with that score exceeds the number of remaining slots; the claimants with that score are randomly assigned to the remaining slots.

The "randomization at the margin" approach used in Kentucky has many positive aspects, including low cost, no direct caseworker involvement and staff perceptions of fairness. Moreover, it provides compelling experimental evidence that addresses the question of the effects of the mandatory reemployment services requirement on claimants at the margin of having it imposed. As the primary policy question in this area concerns small increases or decreases in the budget rather than program termination, this evidence corresponds to the cost–benefit analysis of greatest policy interest.

Perez-Johnson et al. (2011) experimentally evaluate three alternative ways of structuring the "Individual Training Accounts" (ITAs) provided to some participants in the US *Workforce Investment Act* (WIA) program. Everyone receives services but important aspects of the service delivery process differ among the treatment arms. The policy question addressed in this evaluation concerns not keeping or scrapping the WIA program, nor expanding or contracting it, but rather how to operate the ITA component of the program most effectively.

In some cases, elements of service provision are de facto randomized by virtue of the structure of particular programs and researchers can take advantage of these sources of exogenous variation to better understand the associated causal impacts. For example, Oreopoulos (2003) exploits the institutional framework that allocates subsidized housing in Toronto. When they reach the top of the list, families in the queue for subsidized housing are matched to the first appropriately sized accommodation that comes available. These accommodations could be in extremely large, or quite small, housing projects in neighbourhoods of quite different socio-economic status. He uses this "as good as" random assignment to estimate neighbourhood effects on long-term outcomes.

More proactively, the British Columbia government was uncertain about the value of, and need for, annual reviews of its income assistance case files (which both clients and managers consider a burden)—reviews that had been regularly undertaken for years. Warburton and Warburton (2002) discuss how half the caseload was randomly assigned to "no review" as opposed to the standard "annual review." This provided credible evidence for decision-making and is a rare example of an experiment whose operation reduced expenditures, while the cost of analysis was trivial. Other variants of random assignment include randomized rollout of programs too big to put in place in all locations at the same time and randomized encouragement designs, as in Hirano et al. (2000), that randomly assign not treatment but an incentive to participate in the treatment for voluntary programs.

In short, given the tremendous variety of possible randomized designs, we can hardly over-emphasize the potential to use persuasive yet inexpensive (and relatively uncontroversial) experimental evaluations to generate knowledge about program operations and impact.

3.1.4. Ethics, politics and experiments

Policymakers and other stakeholders sometimes express ethical concerns with the random service denial inherent in random assignment designs with "no treatment" or even "less treatment" control groups. Advocates of experiments can address such concerns directly. First, evaluation efforts should focus on programs whose impacts and cost–benefit performance remain uncertain. In such cases, there is no way to tell in advance whether the control group is being randomly punished through denial of valuable services or randomly saved from an ineffective treatment. Second, experimental participants can always be compensated for contributing to the public good of knowledge creation. Unlike the case of medical treatments where larger payments are sometimes made, modest payments should quell any ethical concerns in the social policy domain. Third, an alternative and perhaps weightier ethical concern militates in favour of random assignment. Given limited resources, the operation of programs that do not beneficially impact clients implies the withholding of funds from programs that do have beneficial impacts. How can society ethically allocate taxpayer funds across alternatives without a compelling evidentiary basis when they can easily bring about the production of such evidence?

3.2. Non-experimental evaluation: Selection on observed variables

Now consider the very different case where non-random selection into treatment occurs but the analyst observes all the variables with important effects on both participation and the outcome of interest in the absence of participation. Economists call this case "selection on observed variables" while statisticians call it "unconfoundedness." In this case, the analyst does not observe Z_i; instead it lurks in the shadows producing exogenous variation in treatment status conditional on the observed variables X_i. It implies that, conditional on observed variables, T_i is "as good as" randomly assigned. Formally, this strategy builds on the Conditional (mean) Independence Assumption (CIA), $E(Y_0|X, T=1) = E(Y_0|X, T=0)$.[10]

The CIA represents a very strong assumption indeed! In our view, most evaluations that rely on this assumption fall far short of making a compelling case for it, sometimes because of data limitations and sometimes, more broadly, because we simply lack the knowledge in many policy contexts regarding the variables on which we should condition. Successful application of this strategy requires careful thought about the institutions and economics of the situation in order to make the case that all of the variables that theory and existing empirical knowledge suggest should appear among the conditioning variables in fact do so. Commentators including Heckman et al. (1999, section 6.6) and Rubin (2008) stress this issue. Making this case requires much more than just saying, as many evaluations do, that the evaluation uses "rich" data containing a large number of variables.

10 In the standard notation of the parametric linear model, the CIA becomes $E(u|T, X) = E(u|X)$, where u is the "error" term. The CIA is weaker than the usual exogeneity assumption $E(u|T, X) = 0$. Under the CIA, the conditioning variables justify a causal interpretation of the effect of T but lack any causal interpretation of their own.

It is not the number of conditioning variables that matters, but rather having the ones that make the CIA plausible.

Conversely, conditioning on variables that do not satisfy the CIA, such as intermediate outcomes (a.k.a. "mediators") is equally troubling. For example, conditioning on current occupation when estimating the impact of a labour market program on earnings yields biased estimates because the program may affect earnings via occupational choice. Sometimes researchers even fail to heed the warnings of Bhattacharya and Vogt (2012) and condition on instruments (i.e., variables that affect outcomes only via their effect on treatment choice—more on these shortly).

Policymakers and evaluators can take steps to make the evidence provided by evaluations based on the CIA more compelling. The design of the program can include explicit guidance regarding the factors that gatekeepers should use in making access decisions, which serves to clarify important conditioning variables. Collecting data on factors that are measureable but often go unmeasured, such as attitudes toward work, future orientation (i.e., the subjective discount rate), risk aversion, motivation, social and other non-cognitive skills and the cognitive ability of potential program participants, will also make the selection on observed variables assumption more credible in particular contexts.

An interesting example of the thoughtful choice of conditioning variables comes from a study looking at Ontario physicians' reactions to the offer from the provincial government to shift from fee-for-service to a capitation (fee-per-patient) payment model by Kantarevic and Kralj (2013). In a propensity score matching framework, one of the variables they match on measures the difference in annual earnings if each physician's pre-policy-change pattern of medical practice is priced using both the old and new payment models. That is, it identifies the practice's revenue gain (or loss) from switching to the new payment model holding constant the pre-choice list of billable tasks. Without conditioning on this type of variable (jointly with demographics), selection on observed variables would lack credibility, casting in doubt any causal interpretation. Of course, while including a measure of the revenue change associated with treatment will convince many readers, others might worry that this study lacks explicit conditioning variables capturing the administrative or physic costs (or gains) of switching payment models. The authors make no argument on this front.

In some contexts the literature clearly documents the value of flexibly conditioning on past outcomes measured at a relatively fine level of temporal detail. These pre-participation-choice outcomes implicitly capture many otherwise unobserved variables affecting both treatment choice and outcomes such as motivation, ability and appearance. In the specific context of active labour market programs, see, for example, Heckman, Ichimura, et al. (1998) and Dolton and Smith (2011) on this point.

More broadly, researchers can pursue an agenda that seeks to cumulate knowledge regarding the variables that matter for eliminating selection bias in particular substantive contexts. That agenda can include both within-study

comparisons using experiments as benchmarks—i.e., the literature starting with LaLonde (1986)—as well as the collection and use of new sorts of variables in evaluations to see if they affect the estimates, as with the firm characteristics in Andersson et al. (2013) and the psychological variables in Caliendo et al. (2014). Pursuing such an agenda represents a proactive alternative to the more common strategy of endless carping about potential omitted variables in evaluations that rely on the CIA.

Finally, in cases where the selection on observed variables assumption lacks credibility but where existing knowledge allows for a somewhat informative prior about the nature and extent of the remaining selection bias, formal sensitivity analyses along the lines of those in Altonji et al. (2005) and Koremenos and Smith (2015) for parametric linear models and Ichino et al. (2008) for matching estimators can indicate the substantive and inferential consequences of reasonable departures from the CIA. We think the literature would benefit from more analyses along these lines.

When relying on the CIA, analysts typically estimate a parametric linear regression model by least squares, or a non-linear parametric model such as a probit by maximum likelihood, or employ matching or weighting estimators based on the estimated propensity score, $\Pr(T = 1|X)$. In general, weighting and matching estimators represent the first choice for various technical reasons, provided the sample size justifies their use.[11] Angrist (1998) points out that, in a world of heterogeneous treatment effects, the OLS estimand in the parametric linear regression model does not correspond to the ATET, while the estimand associated with commonly used matching and weighting estimators does correspond to the ATET. The difference arises from the fact that OLS and matching weight the data differently.[12]

Aside from relaxing functional form assumptions, a major advantage of matching estimators is that they push researchers to think carefully about common support issues. In the region of common support, the probability of participation is bounded away from zero and one, i.e., $0 < \Pr(T = 1|X) < 1$. Most matching estimators, properly applied, produce estimates only on the region of common support. As noted in Black and Smith (2004) and Crump et al. (2009), limiting estimation to the common support changes the sub-population to which the treatment effect applies. At the same time, it may substantially increase the credibility of the reported estimates. In contrast, the parametric linear model happily ignores support issues by implicitly projecting impacts into regions with only treated or only untreated units (or neither). While researchers could examine support issues when estimating parametric linear models, they rarely

11 See Huber et al. (2013), Busso et al. (2014) and Frölich et al. (2015) for technical details and Monte Carlo comparisons of alternative estimators.

12 Another class of estimators combines matching or weighting with parametric linear regression. This class includes the widely used (outside of economics) "double robust" estimator that combines inverse propensity weighting with parametric linear regression. See, for example, Ho et al. (2007), the studies cited in the preceding footnote and the discussion in Imbens and Wooldridge (2009, section 5.8).

do so in practice. As such, we think matching adds value, even if only as a supplementary feature of an analysis, by letting researchers know whether or not they have a support problem. They can then, if they choose to, make an explicit case for extrapolating impacts beyond the region of common support.[13]

3.3. Instrumental variables

Instrumental variables (IV) sometimes provide consistent estimates of causal parameters in contexts where the CIA does not hold given the available data.[14] Not surprisingly, the application of instrumental variables requires a credible instrument (a version of Z from section 2), which need not exist in many contexts. A bit loosely, an instrument is a variable that: (i) affects participation in the treatment sufficiently strongly but (ii) is conditionally uncorrelated with outcomes other than through its effect on participation. The literature calls (i) the "first stage" condition (in reference to the common application of two-stage least squares when using instrumental variables) and (ii) the "exogeneity" or "exclusion restriction" condition. These two properties sometimes conflict in practice as many naturally occurring candidate instruments either strongly predict treatment but lack credible exogeneity or appear credibly exogenous but only weakly predict treatment.

Where do good instruments come from? Sometimes nature provides instruments, which can be as diverse as fluctuations in rainfall or temperature or the sex composition of children. Social events, governments and other institutions sometimes provide exogenous variation in forms such as strikes, changes in compulsory schooling laws, or even random fluctuations in emergency room admissions. Often, this variation represents plausibly exogenous variation in the cost of treatment, C_i. In each situation, researchers need to make a good case that their instrument has the properties of a valid instrument described above, or else abandon the effort. Contrary to what one might casually infer from reading the applied literature, we know of no systematic evidence that instrument validity increases in instrument cleverness.

In thinking about how to evaluate candidate instruments, we begin with the first stage condition (i). Following the literature, we focus on binary instruments, i.e. $Z \in \{0, 1\}$, with some remarks at the end about the more general case. In the binary IV case, the first stage condition becomes $\Pr(T = 1|Z = 0) \neq \Pr(T = 1|Z = 1)$. IV is consistent, but biased in finite samples. The extent of the finite sample bias (which is toward the corresponding OLS estimate) depends on the strength of the first stage relationship. Bound et al. (1995) show that substantial bias can emerge even in very large samples in the absence of a sufficiently strong first stage relationship. All applications of instrumental variables should consider the strength

13 See, for example, Heckman, Ichimura, et al. (1998), Angrist (1998), Smith and Todd (2005) and Imbens (2015) for further methodological discussions.

14 Both difference-in-differences and regression discontinuity represent special cases of instrumental variables; we follow the literature in discussing them separately.

of the first stage in light of the various results in the "weak instruments" literature, which includes the famous rule of thumb that the first stage F-statistic on the instruments (not the entire first stage) should exceed 10 offered up in Stock et al. (2002).[15]

The second requirement, condition (ii), embodies two conceptually distinct parts. First, a valid instrument may not have a direct effect on the outcome variable of interest. Second, a valid instrument may not have an indirect effect operating through a channel other than the treatment under study. In a parametric linear common effect world, i.e., a world in which the treatment has the same effect on all units, these two conditions correspond to (assuming) a zero correlation between the instrument Z and the outcome equation error (e.g., the error term in the second stage of two stage least squares).

To put some substantive flesh on these conceptual bones, consider an example with earnings as the outcome variable of interest, high school completion as the binary treatment and the mandatory school leaving age as the binary instrument, as in Oreopoulos (2006), where, for simplicity, we imagine that some jurisdictions in some periods have a school leaving age of 16 ($Z = 0$) and others a school leaving age of 17 ($Z = 1$). The first conceptual condition of requirement (ii) necessitates that the direct effect of the school leaving age operates only through years of schooling. If a higher school leaving age also induces students to, say, work harder and learn more conditional on their years of completed schooling then this condition fails. The second conceptual condition in (ii) requires that a higher school leaving age does not correlate with other policies (for example, higher quality vocational education) or other variables not included in the conditioning set, such as parental education, that also increase earnings conditional on high school completion and the included covariates.

Another way to think about what instruments do sees them as extracting a particular subset of the variation in the endogenous variable. If we mentally divide the variation in the endogenous variable into "good" (not correlated with the outcome equation error term) and "bad" (correlated with the outcome equation error term) variation then instruments isolate a subset of the "good" variation and use only that variation in estimation. The cost of throwing out the "bad" variation (and, typically, much of the "good" variation as well) comes in the form of larger standard errors, reflecting the reduction in effective variation in the treatment. While economists typically treat consistency and variance lexicographically, with variance minimized conditional on finding a consistent estimator, a mean squared error criterion would sometimes prefer inconsistent but more precise OLS estimates to consistent but imprecise IV estimates, as noted in Black et al. (2015).

15 In our experience, graduate students have an uncanny knack for uncovering instruments with first-stage F-statistics between 9 and 10. See Murray (2006) for an accessible introduction to the weak instrument literature. Angrist and Pischke (2009) suggest alternatives to two-stage least squares for instruments of marginal strength.

Moving to a heterogeneous treatment effects world (i.e., a realistic world) opens up the possibility of a second type of selection bias. The traditional instrumental variables approach (and, indeed, the earlier applied literature more broadly) worried only about non-random selection into treatment based on unobserved determinants of the outcome, as with the literature on ability bias in the estimation of the effects of schooling. In contrast, the modern approach embodied in the model in section 2 worries about selection both on the unobserved component of the untreated outcome and on the idiosyncratic component of the impact. For example, holding costs constant, we expect students choosing between finishing high school and dropping out to select into finishing high school when they expect worse labour market outcomes without finishing (i.e., a smaller opportunity cost) and higher payoffs from high school completion (i.e., a larger treatment effect).

In the heterogeneous treatment effects world the outcome equation error term includes both the unobserved component of the untreated outcome and, for the treated units, the idiosyncratic component of the unit-specific impact. In order to interpret the IV estimand as the ATET, the instrument must have a zero correlation with both of these components. The literature on applied econometrics has spent the last decade or so coming to grips with the fact that many common classes of instruments likely fail this condition.

To see this, return to the case of an experiment with imperfect compliance introduced in section 3.1. As noted by Heckman (1996a), random assignment represents a special case of instrumental variables. Here $Z = 1$ denotes random assignment to the treatment group and $Z = 0$ random assignment to the control group. Now suppose that the randomized population consists of individuals who know their impact of treatment and that for one third it equals 200, for one third it equals 100 and for one third it equals zero. Suppose further that participating in treatment costs 50 in the treatment group and, because of the necessity of finding an alternative provider, 150 in the control group. In a simple maximizing model treatment group members with impacts of 200 and 100 take the treatment, while those with impacts of zero do not, as it fails a cost–benefit test for them. Similarly, in the control group only, those with impacts of 200 take the treatment, as it only passes a cost–benefit test for them. Thus $\Pr(T = 1|Z = 1) = 2/3$ and $\Pr(T = 1|Z = 0) = 1/3$. The now-standard terminology of Imbens and Angrist (1994) calls those with an impact of 200 "always takers," because they take the treatment for both values of the instrument (i.e., when assigned to either the treatment group or the control group). Similarly, it calls those with a zero impact "never takers." The term "compliers" captures those who take treatment when assigned to the treatment group but not when assigned to the control group, which is to say that they comply with the intent of the experimenter.

With imperfect compliance, $Z \neq T$ but so long as $\Pr(T = 1|Z = 1) \neq \Pr(T = 1|Z = 0)$, Z will continue to satisfy the "first stage" condition. Thinking about assignment to the control group as raising the cost of receiving treatment makes it natural to assume, as in our example, that anyone who receives treatment when assigned to the control group would also do so when assigned to the treatment

group. This represents an application of what the Imbens and Angrist (1994) call the "monotonicity" assumption. Monotonicity rules out individuals who defy the intent of the experiment by reversing the behaviour of the compliers. In this example, these "defiers," as the literature calls them, would take treatment when assigned to the control group but not when assigned to the treatment group.

Now think about the simple IV estimator in the context of our contaminated experiment. It divides the experimental mean difference by the difference in treatment probabilities. This yields a value of 100 because the impact on the always-takers cancels out in the experimental mean difference. In general, as in this example, in a heterogeneous treatment effects world, and assuming monotonicity, IV estimates the mean impact on compliers, which the literature calls the Local Average Treatment Effect (LATE).

Returning to our earlier example, we see that changes in the compulsory schooling age induce variation in schooling levels only for a particular subset of the population. For example, increasing the age from 15 to 16 years in the Canadian institutional context will affect only those individuals contemplating dropping out prior to high school completion. The resulting treatment effect of additional schooling refers only to those individuals whose schooling changes as a result of the policy change and not to individuals who would graduate and go to university (or drop out at 14 or 15) regardless of the value of the compulsory schooling age.

In general, the literature provides only a very limited array of strategies for testing the exogeneity assumption that underlies a valid instrument. In a common effect world, classical Durbin–Wu–Hausman tests provide some guidance via tests of the equality of estimates based on different instruments. This strategy falls apart in the heterogeneous treatment effects world where each instrument identifies its own distinct LATE. Some substantive contexts admit "placebo" (or falsification) tests of over-identifying restrictions. In these tests, the instrument gets applied to a separate but related context (e.g., a different time period or different jurisdiction) where the LATE is known to equal zero. Finally, one can compare the LATE estimate from a candidate instrument to an experiment designed to estimate a LATE, as in Black et al. (2016), but such experiments remain few in number.

Instead of testing, analysts must generally make the case for the instrument using the relevant theory, along with information about the institutional context and prior knowledge regarding the determinants of treatment and outcomes (and referees, editors and readers must assess those arguments). In the case of cost-based instruments, the simplest of theory motivates both the first stage and monotonicity. Institutional knowledge often helps to rule out competing pathways from the instrument to the outcome. And, of course, covariates matter here too. Adding particular conditioning variables may make the exogeneity assumption more plausible for particular instruments, as with our example of parental education levels in the context of the mandatory schooling age. This process of argumentation renders many instrumental variable estimates quite controversial.

That different instrumental variables identify LATEs corresponding to different complier groups complicates interpretation both within and across studies. Hertzman et al. (2014) explore child apprehension rates in the British Columbia foster care system using two quite different instruments and find that the LATE identified via a measure of caseworker-specific administrative discretion in a system operating "normally" differs from the LATE arising from an external shock to the foster care system that simultaneously increased apprehension rates across all caseworkers (and thereby changed the system, altering its scale, so that at least in the short run it was not operating normally).

In common with Doyle (2008), to justify the monotonicity condition, and thereby the LATE interpretation when using caseworker discretion as an instrument, these authors assume that lenient and strict social workers rank the risks children face in a similar manner but differ regarding the threshold for apprehension. In contrast, if social workers disagree on the ranking, with low-probability-of-apprehension social workers taking children into care that higher-probability-of-apprehension social workers would consider safe with their family, then the monotonicity assumption fails. In this case, the IV estimand represents not a LATE but rather a "muddle" unless, as pointed out by Angrist et al. (1996), the compliers and defiers happen to have the same average treatment effect.

Some instruments identify LATEs of great relevance to policy, while others do not. Usually instruments will not identify the ATET parameter, which means that IV estimates typically cannot directly answer the "keep it or cut it" question that underlies many cost–benefit analyses. On the other hand, an instrument that varies, say, the costs of program participation at the margin, may provide exactly the parameter of interest if the policy change under consideration consists of modest spending increases to reduce the costs of program access. See, for example, Angrist and Fernández-Val (2014) for more on generalizing and comparing LATEs.

The Marginal Treatment Effect (MTE) framework of Heckman and Vytlacil (2005) generalizes the standard LATE setup to encompass discrete and continuous instruments and other averages of heterogeneous treatment effects while maintaining the monotonicity assumption. The MTE is the treatment effect for units at the margin of treatment given a particular value of a particular instrument. The LATE, the ATET and the ATE all then constitute particular integrals over the distribution of MTEs. In practice, the set of treatment probabilities generated by a given instrument limits the set of MTE values over which integrals can be calculated. For an application see Doyle (2008) or Heckman et al. (2011), who emphasize that "local" as defined by a particular instrument may not be "marginal" for some particular policy question.[16]

16 Chapter 4 of Angrist and Pischke (2009) provides a good introduction to IVs. Lewbel et al. (2012) survey the binary choice context. Heckman et al. (2001) offer a broad conceptual framework for thinking about instruments. Heckman and Urzúa (2010) discuss the limitations of instrumental variables methods.

3.4. Longitudinal methods

Longitudinal methods use variation over time in treatment status to estimate the impact of treatment. Credible application of longitudinal methods relies on a clear understanding of the process by which some units come to receive treatment at particular times while other units receive treatment at other times or not at all. In our experience, most applied papers using longitudinal methods and employing causal language make no explicit case for why the reader should believe the assumptions required for a causal interpretation. To see the importance of understanding the treatment assignment process, consider, for example, the large literature on earnings changes with job changes (e.g., Morissette et al. 2007). Differencing across voluntary quits implies quite a different (and non-causal) interpretation compared to differencing across job changes resulting from (involuntary-to-the-worker) plant or firm closures. Even for the latter group differences in pre-separation earnings trajectories and the opportunity for workers to quit prior to a plant closing, represent potential sources of bias.

The simplest longitudinal method compares the outcomes of treated units before and after treatment with no comparison group. Researchers may apply this "before–after estimator" (or "interrupted time series design" for those with higher consulting rates) to individuals, as when comparing outcomes before and after participation in a training program, or to jurisdictions, as when comparing alcohol-related fatalities at the province level before and after a change in the minimum legal drinking age. Such before–after comparisons implicitly assume that in the absence of the treatment or policy change, expected outcomes in the "after" period would have equalled expected outcomes in the "before" period. Sometimes this assumption makes sense and other times it does not. It fails when other factors affecting outcomes also change over time.

Concerns about the plausibility of simple before–after comparisons have led many researchers to prefer the "difference-in-differences" (DiD) estimator, which introduces a comparison group; see Card et al. (2011) for a practical guide and contrast to randomization. This estimator compares the before–after change in outcomes of the treated units to the before–after change in the outcomes of a sample of untreated units. DiD is a special case of a more general class of panel data estimators that rely on within-unit variation over time to estimate the impacts of programs or policies, using untreated units to control for counterfactual trends in outcomes for the treated units. Both DiD and more general panel data studies rely on the "bias stability" or "common trend" assumption which holds that the before–after change in outcomes for the treated units would, in the absence of the program or policy, equal (at least in expectation) that for the untreated units.[17] Put differently, any differences between $E(Y_0|T=1)$ and $E(Y_0|T=0)$, or their

17 A more general assumption that allows selection into treatment based on linear time trends in the untreated outcome rather than just on time invariant differences in outcome levels underlies the so-called random growth model. Identification requires at least two periods of pre-treatment outcomes. See, for example, Moffitt (1991).

conditional on covariate analogues, must remain constant over time. Some parts of the literature refer to this situation (perhaps a bit misleadingly) as a "natural experiment"; for further discussion, see Meyer (1995).

In certain contexts, the bias stability assumption will make sense when an assumption of no change in expected outcomes in the absence of treatment for the treated units would not. At the same time, DiD is not a panacea. Researchers need to make a solid case for the assumption's reasonableness in their context and readers need to judge the plausibility of those arguments to determine the credibility of the causal claims. For example, selection into treatment based on transitory outcome shocks implies failure of the common trends assumption. Thus, much of the intellectual action when using these methods centers on *how* and *when* the treated units came to be treated (and why the comparison units were not treated). Analysts must also worry about anticipatory changes in behaviour prior to a treatment actually starting but as a direct result of its impending arrival, as when customers rush to buy prior to a sales tax increase.

Some examples will clarify these issues and also illustrate the many different types of comparison groups employed within this estimation framework. Heckman and Smith (1999) examine DiD in the context of a job training program. The comparison group consists of eligible non-participants in the same local labour markets as the participants. Using an experimental benchmark, they find that DiD performs poorly in their context, exhibiting both bias and strong sensitivity to the choice of before and after periods. This poor performance results from the fact that training program participants select (in part) into training based on transitory labour market shocks—typically job loss.

The famous minimum wage paper of Card and Krueger (1994) illustrates DiD applied at the jurisdictional level. Their paper, as well as the companion paper by Neumark and Wascher (2000) that uses (arguably) better data and obtains a somewhat different answer, compares the changes in employment in a set of fast food restaurants in a local labour market that straddles the New Jersey and Pennsylvania border before and after an increase in the minimum wage that affects only New Jersey. The focus on a single labour market plays a key role in the plausibility of the estimates, though it also raises the possibility of spillover effects. Milligan and Stabile's (2007) evaluation of changes to Canada's National Child Benefit using DiD across provinces provides another example using jurisdictional policy variation.

Many longitudinal studies do little to convince the reader that the timing of treatment does not depend on transitory changes in the outcomes of interest. In some cases, justifying a source of variation as exogenous will be easier when it emanates from a third party, such as jurisdiction-level policy changes affecting individuals, than when it is tied up with individual choices. Both the data at hand and institutional knowledge can reveal the importance of selection into treatment based on transitory outcome shocks. DiNardo and Lee (2010) emphasize the value of "falsification" (or "placebo" or "pre-program") tests based on impact

estimates in periods prior to treatment when the true value equals zero under bias stability. When rolling out new programs, governments can deliberately stagger the rollout in ways unrelated to untreated outcomes so as to allow a compelling causal analysis using longitudinal methods.[18]

3.5. Regression discontinuity

Regression discontinuity (RD) designs exploit discontinuous changes in treatment receipt that typically result from discontinuities in program rules. The RD design has the great virtue of conceptual simplicity. In situations where assignment to treatment depends on a continuous variable, such as a test score or proposal rating, and where the probability of treatment changes abruptly at a particular cut-off value of that continuous variable, a comparison of mean outcomes just above and just below the cut-off can provide a compelling source of information about treatment effects. The literature calls the continuous variable that determines treatment assignment the "running variable." The econometric literature defines a number of different estimators for the RD case, but they all represent different ways of taking (weighted) averages of outcomes on the two sides of the discontinuity.

In thinking about exactly what treatment effect gets estimated in the context of a particular discontinuity, it helps to distinguish between "sharp" and "fuzzy" RD designs (i.e., between Z perfectly or probabilistically determining T as in section 3.3.1). In a sharp design, the probability of treatment moves from zero to one at the cut-off value. In this case, RD identifies the average treatment effect for units whose characteristics put them *at the discontinuity*. In a fuzzy design, the probability of treatment need not equal zero or one on either side of the cut-off but it must change discontinuously at the cut-off. For example, publicly funded flu shots for those over age 65 could induce a discontinuity in the probability of receiving a flu shot at that age. In the fuzzy case, under certain pesky but often plausible additional assumptions, the RD design identifies the LATE on those units that change their treatment status at the cut-off. For example, in the case of the flu shots, a comparison of health outcomes on either side of the cut-off at age 65 would yield the mean impact of receiving a flu shot on individuals aged exactly 65 who would not get a shot unless it were free, the "compliers" in the language of IV. It does not provide information on the impact of a shot on those near the cut-off who would, or would not, get one irrespective of price or on individuals much younger or older than 65.[19]

In both the sharp and fuzzy cases, generalizing the estimated impacts to units with values of the running variable other than the value at the cut-off requires

18 Longitudinal estimators raise particular applied econometric issues. Bertrand et al. (2004) highlight issues related to serial correlation of the error terms. Cameron and Miller (2015) and MacKinnon and Webb (2016) address issues related to small numbers of cross-sectional units. For more on longitudinal methods in a treatment effects context see Lechner (2010). Heckman (1996b) critiques the application of DiD methods.

19 As in section 3.3, monotonicity rules out individuals who become less likely to get a flu shot when the price falls.

additional assumptions. The plausibility of such assumptions depends on prior knowledge and the institutional context; see, for example, Wing and Cook (2013) for an extended discussion.

Lemieux and Milligan (2008) utilize a sharp regression discontinuity design to examine a policy change in Quebec that eliminated lower social assistance benefits for childless recipients under age 30 compared to those over age 30. In the presence of a large policy change (a 175% change in benefits at the cut-off) they find strong evidence that the increased generosity of social assistance benefits modestly reduced employment.

RD, like other identification strategies, has its issues. Using either calendar time or age to define a discontinuity raises the potential for bias due to anticipatory behaviour. RD requires sufficient data near the discontinuity to estimate a treatment effect with reasonable power—a sometimes difficult standard to reach, as documented in Schochet (2008). The discontinuity must build on a running variable that both the program and the evaluators can measure without much error and that potential participants or program staff cannot easily manipulate in order to change their treatment status. For example, a generous subsidy to firms with 10 or fewer employees will induce some firms to change their number of employees from 11, 12 or 13 down to 10 in order to qualify for the subsidy. Such behaviour invalidates the regression discontinuity design, as the firms on one side of the margin (with 10 employees) no longer look like the firms on the other side of the margin (with 11 employees) due to the self-selection. More broadly, the substantive case for a causal interpretation of an RD estimand typically relies on detailed institutional knowledge of the assignment process along with formal testing, as in the McCrary (2008) density test for potential manipulation and commonly used tests of covariate balance at the discontinuity.[20]

Finally, the opportunity to estimate impacts using RD methods depends almost entirely on program design decisions made by policymakers and program managers. Many of the existing evaluations using RD methods rely on the "luck" of having available institutions that happen to embody useful discontinuities. Policymakers and program operators should think prospectively about how to design programs to embody discontinuities that will yield useful impact estimates.

4. Reductionist enthusiasms: Hierarchies and magic bullets

Both inside economics, as in Leigh (2009), and outside economics, as in Guyatt et al. (2008), in the health/epidemiology literature, one sometimes sees proposals related to hierarchies of evidence that rank alternative causal identification strategies. Typically, random assignment (the "gold standard") tops the list, followed by discontinuity designs, instrumental variables and bias stability (e.g.,

20 Cook (2008) gives a broad history of regression discontinuity in the social sciences. For methodological details see the fine surveys by van der Klaauw (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010).

difference-in-differences) designs, then studies relying on selection on observed variables and, finally, before–after comparisons (with, occasionally, case studies, theory or expert opinion (!) rounding out the field).

We do not dispute that if one did a serious, impartial, quality ranking according to well-defined and generally agreed-upon criteria that the average quality of well-executed evaluation studies using each method would likely correspond to such an ordering. At the same time, we worry that hierarchies, in their rush to reduce the amount of troublesome thinking required in evaluating evidence, focus solely on the between-strategy variation in study quality while ignoring the (quite substantial) within-strategy variation. As such, at the margin, they encourage weak papers using, say, random assignment and RD and discourage strong papers using, say, selection on observed variables methods. Sometimes, given a particular dataset and context, nominally moving "up" the hierarchy may make things worse rather than better, as when a weak and/or not obviously valid instrument replaces a reasonably compelling set of conditioning variables, thereby inflating the standard errors and quite possibly increasing the bias as well. In sum, we think of hierarchies of evidence as useful for governments and others in considering the design of a research program, but applied simplistically, they represent a flawed and ultimately counter-productive substitute for thinking that attempts to institutionalize the credibility revolution into a ritualistic identification strategy choice algorithm.

A second enthusiasm leads directly to the (in this sense) misguided literature set in motion by LaLonde (1986). It seeks the holy grail of non-experimental evaluation: a method that always and everywhere solves the selection problem. Dehejia and Wahba (1999, 2002) represent famous papers in this approach, which many (not including their authors) have interpreted as showing that matching "works" in the sense of always solving the selection problem. These papers spawned a large literature addressing the question "does matching work?" by comparing matching estimates to experimental estimates, sometimes using remarkably weak sets of conditioning variables in the matching. In fact, the question "does matching work?" is ill posed. Matching "works" in the sense of providing consistent estimates when the available variables suffice for the CIA to hold, and not otherwise. Thus, we know the answer to the generic "does matching work?" question in advance; it is "sometimes, but only when the data and context support it."

Rather than searching for a non-existent magic bullet estimator, we argue that the literature should seek to build a body of knowledge on what combinations of identification strategy and data work for particular combinations of institutions and questions of interest. Rather than relying on a hierarchy or on this year's magic bullet to choose an identification strategy for a particular project, researchers should seek to use the strategy best suited to providing a compelling impact estimate. And, of course, credibility includes an acknowledgement of the limits on what can be claimed with confidence. Sometimes no design-based econometric path to credible causal evidence exists given the institutional context and available data.

5. Spillovers and other general equilibrium effects

General equilibrium effects of programs and policies include both their effects on persons, organizations or markets that do not directly participate and spillover effects among participants. Such effects have proven, in general, quite difficult to pin down, but we argue that, contrary to the belief implicit in much of the literature, "difficult to estimate" does not imply "equals zero."

Evaluations can often pick up direct spillovers via thoughtful data collection. For example, an educational intervention increasing the amount of classroom time devoted to mathematics in primary school should collect outcome data not only on math achievement but also on achievement in the subjects whose classroom time gets reduced. Evaluations of labour market programs should collect data on social outcomes such as volunteering, criminal behaviour and child outcomes and behaviour, as in the Morris and Michalopoulos (2003) analysis of Canada's Self-Sufficiency Project.

Evaluators sometimes obtain estimates of spillovers by assigning treatment at the group (or location) level and measuring outcomes for both participants and non-participants. For example, the clever village-level random assignment in the evaluation of the PROGRESA conditional cash transfer program in Mexico, combined with the collection of data on both eligible and ineligible households in both treatment and control villages allows Angelucci and De Giorgi (2009) to provide a subtle analysis of within-village spillovers from the program. Crépon et al. (2013) represents an apogee of this approach. In their study of a French active labour market program, they randomly assign both eligible individuals within jurisdictions and the fraction of the eligible population treated at the jurisdictional level. Using the cross-jurisdiction variation in the experimental impacts as a function of the fraction of the eligible treated they pin down substantively important effects on non-participants.

In many cases, obtaining estimates of general equilibrium effects will require writing down and either estimating or calibrating a structural model of the relevant market. This approach represents a major investment of evaluator time and energy and requires a different skill set, more like that of modern macroeconomics, than that possessed by many in the partial equilibrium causal impact (i.e., program evaluation) business. Most evaluations should, however, draw on this broader literature when discussing the nature and extent of equilibrium effects in particular contexts and their potential to affect the conclusion of a cost–benefit analysis.

Three examples highlight the power of this sort of analysis, along with its effort costs and heavy reliance on economic theory in general and specific functional form assumptions in particular. Davidson and Woodbury (1993) look for displacement effects in one of the US Unemployment Insurance (UI) bonus experiments, which paid claimants a lump sum if they ended their claim within a specific number of weeks. They estimate that the displacement of workers not in the experiment cancelled out about 20% of the employment impact of the program

estimated in the experiment. In a study of tuition subsidy programs for university students, Heckman, Lochner et al. (1998) find much larger general equilibrium effects. In their study, the partial equilibrium estimate of the impact of treatment on the treated is 10 times larger than a general equilibrium impact that accounts for the decline in the relative wage of persons with a university degree resulting from an increased supply. Finally, Lise et al. (2004) examine the general equilibrium effects of Canada's Self-Sufficiency Project. They find that taking account of the program's effects on the job search behaviour of other workers (and of the single parents themselves early in their spells of income assistance receipt) leads to a reversal of the positive cost–benefit conclusions reached in the partial equilibrium experimental evaluation.

6. Cost–benefit analyses

Cost–benefit analysis exposes the full range of costs and benefits associated with a policy or program by requiring their itemization, justification and valuation. For reasons of time and space, we do not attempt a full consideration of cost–benefit analysis. Instead, we highlight a small number of important issues often ignored in practice.

First, we emphasize the importance of doing a full-blown cost–benefit analysis, especially for large and/or expensive programs and particularly influential ones such as Perry Preschool (e.g., Heckman et al. 2010). Second, we highlight the importance of considering multiple outcomes. For example, employment and training programs may have impacts on outcomes other than earnings and employment, such as participation in transfer programs, health, marital and family behaviour, volunteering and crime. Some outcomes present real challenges to the analyst who must convert them to dollar terms, as with primary school test scores. But, as noted in relation to general equilibrium effects, "hard to measure" does not imply "equals zero." Third, evaluations of labour market programs (in particular) need to account for the lost value of participant "leisure" (which may include childcare and the care of sick and aging relatives), as emphasized in Greenberg and Robins (2008).

Fourth, a complete analysis should account for the marginal social cost of public funds (a.k.a. the "excess burden") and so recognize that a dollar of public funds costs society more than a dollar due to distortionary taxation (Dahlby 2008). Fifth, evaluations typically have available only a few years of follow-up data. For programs expected to have impacts in the long term, this implies projecting the impact estimates to time periods outside the data. In some cases, the cost–benefit performance of a program may depend critically on these projections. We suggest presenting the results of the cost–benefit analysis conditional on multiple assumptions about the persistence of any estimated program impacts, as in Andersson et al. (2013). The assumptions about impact persistence should build on findings on the persistence of impacts in similar programs drawn from

the literature. Sixth, most programs incur costs in the short term but reap their benefits, if any, in both the present and the future. Taking proper account of the timing of benefits requires the discounting of future benefits (and costs, if any) back to the present. Doing this, in turn, requires a well-justified social discount rate, as discussed by Burgess (2010).

Seventh, we often experience a sense of wonder when we learn, in response to questions about the cost of particular public programs, that no good information exists. Serious cost–benefit analysis requires good data on marginal and average costs, data that public agencies ought to have handy in any event to guide their decision-making. Finally, a complete cost–benefit analysis should take account of general equilibrium effects when possible. This may require a separate evaluation component or it may rely on estimates from the literature for similar programs.[21]

7. Concluding remarks

As a discipline, economics' relevance depends in large part upon producing ideas (theories) and empirical results useful to, and perceived as credible by, the broader society. Hard-headed estimates of causal impacts as well as descriptive analyses, interpreted in light of the relevant economic theory, represent fundamental contributions to evidence-based policy.

We reject all substitutes for thinking seriously about the choice of applied econometric method in light of the available data, the institutional context and the policy question of interest. These substitutes include both hierarchies of evidence and related magic bullet theories regarding the universal superiority of particular identification strategies or estimators. They also include low-grade substitutes (at least as currently operationalized) such as the performance management and participant evaluation approaches discussed in Smith and Sweetman (2010).

Instead, we advocate for the careful selection of an applied econometric approach in light of the relevant economic theory, the available data, the institutional context and the policy and/or academic question of interest. The realization among applied researchers of the importance of heterogeneous treatment effects and of selection by agents (and by gatekeepers) into and out of programs and policies based on beliefs about these heterogeneous impacts has complicated this task. Different identification strategies imply different, and sometimes quite limited, causal estimands; generalizing beyond these narrow interpretations requires analytical care. In discussing the available identification strategies, we have highlighted the ways in which researchers can make the case for a causal interpretation by marshalling knowledge of the context (particularly the institutions governing treatment choice), along with economic theory and statistical testing. In our view, making such explicit cases for causality remains a margin with marvellous

21 See the discussions in Heckman et al. (1999), Smith and Sweetman (2010) and Barnow and Smith (2016) for expanded versions of these arguments.

opportunities for improvement; in many cases, even attempting such a case would represent an important step forward.

Finally, we view thorough cost–benefit analysis as a critical final step in program evaluation. Such analyses represent an important bridge between academic economics and the world of policy, both a contribution by economics to the wider society and an advertisement of the value of our profession.

References

Abbring, J., and J. Heckman (2007) "Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation." In *Handbook of Econometrics*, 6B, pp. 5145–303, eds. J. Heckman and E. Leamer. Amsterdam: Elsevier

Altonji, J., T. Elder, and C. Taber (2005) "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy* 113(1), 151–84

Andersson, F., H. Holzer, J. Lane, D. Rosenblum, and J. Smith (2013) "Does federally-funded job training work? Nonexperimental estimates of WIA training impacts using longitudinal data on workers and firms," NBER working paper no. 19446

Angelucci, M., and G. De Giorgi (2009) "Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption?," *The American Economic Review* 99(1), 486–508

Angrist, J. (1998) "Estimating the labour market impact of voluntary military service using social security data on military applicants," *Econometrica* 66(2), 249–88

Angrist, J., and I. Fernández-Val (2014) "ExtrapoLATE-ing: External validity and overidentification in the LATE framework." In *Advances in Economics and Econometrics*, vol. 3, pp. 401–35, eds. D. Acemoglu, M. Arellano, and E. Dekel. Cambridge University Press

Angrist, J., G. Imbens, and D. Rubin (1996) "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association* 91(434), 444–55

Angrist, J., and J. S. Pischke (2009) *Mostly Harmless Econometrics*. Princeton: Princeton University Press

Banerjee, A., and E. Duflo (2009) "The experimental approach to development economics," *Annual Review of Economics* (1), 151–78

Barnow, B. (2010) "Setting up social experiments: The good, the bad and the ugly," *Journal for Labour Market Research* 43(2), 91–105

Barnow, B., and J. Smith (2016) "Employment and training programs." In *Economics of Means-Tested Transfer Programs in the United States*, vol. 2, forthcoming, ed. R. Moffitt. Chicago: University of Chicago Press for NBER

Bertrand, M., E. Duflo, and S. Mullainathan (2004) "How much should we trust differences-in-differences estimates?," *The Quarterly Journal of Economics* 119(1), 249–75

Bhattacharya, J., and W. Vogt (2012) "Do instrumental variables belong in propensity scores?," *International Journal of Statistics & Economics* 9(A12), 107–27

Björklund, A., and R. Moffitt (1987) "The estimation of wage and welfare gains in self-selection models," *The Review of Economics and Statistics* (69), 42–49

Black, D., J. Galdo, and J. Smith (2016) "Evaluating the regression discontinuity design using experimental data," unpublished manuscript, University of Michigan

Black, D., J. Joo, R. LaLonde, J. Smith, and E. Taylor (2015) "Simple tests for selection bias: Learning more from instrumental variables," IZA discussion paper no. 9346
Black, D., and J. Smith (2004) "How robust is the evidence on the effects of college quality? Evidence from matching," *Journal of Econometrics* 121(1), 99–124
Black, D., J. Smith, M. Berger, and B. Noel (2003) "Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system," *The American Economic Review* 93(4), 1313–27
Blundell, R., L. Dearden, and B. Sianesi (2005) "Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey," *Journal of the Royal Statistical Society*, 168 (Series A, Part 3), 473–512
Bound, J., D. Jaeger, and R. Baker (1995) "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association* 90(430), 443–50
Burgess, D. (2010) "Toward a reconciliation of alternative views on the social discount rate." In *Discount Rates for the Evaluation of Public–Private Partnerships*, pp. 131–56, eds. D. Burgess and G. Jenkins. Montreal: McGill–Queen's University Press
Busso, M., J. DiNardo, and J. McCrary (2014) "New evidence on the finite sample properties of propensity score reweighting and matching estimators," *The Review of Economics and Statistics* 96(5), 885–97
Caliendo, M., R. Mahlstedt, and O. A. Mitnik (2014) "Unobservable, but unimportant? The influence of personality traits (and other usually unobserved variables) for the evaluation of labor market policies," IZA discussion paper no. 8337
Cameron, C., and D. Miller (2015) "A practitioner's guide to cluster-robust inference," *The Journal of Human Resources* 50(2), 317–72
Card, D., and D. Hyslop (2009) "The dynamic effects of an earnings subsidy for long-term welfare recipients: Evidence from the Self-Sufficiency Project applicant experiment," *Journal of Econometrics* 153(1), 1–20
Card, D., P. Ibarrarán, and J. M. Villa (2011) "Building in an evaluation component for active labor market programs: A practitioner's guide," IZA working paper no. 6085
Card, D., and A. Krueger (1994) "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review* 84(4), 772–93
Cook, T. (2008) "Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics," *Journal of Econometrics* 142(2), 636–54
Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013) "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment," *The Quarterly Journal of Economics* 128(2), 531–80
Crump, R., V. J. Hotz, G. Imbens, and O. Mitnik (2009) "Dealing with limited overlap in estimation of average treatment effects," *Biometrika* 96(1), 187–99
Dahlby, B. (2008) *The Marginal Cost of Public Funds: Theory and Applications*. Cambridge, MA: The MIT Press
Davidson, C., and S. Woodbury (1993) "The displacement effects of reemployment bonus programs," *Journal of Labour Economics* 11(4), 575–605
Deaton, A. (2010) "Instruments, randomization, and learning about development," *Journal of Economic Literature* 48(2), 424–55
Dehejia, R., and S. Wahba (1999) "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association* 94(448), 1053–62
——— (2002) "Propensity score matching methods for non-experimental causal studies," *The Review of Economics and Statistics* 84(1), 151–61

DiNardo, J., and D. Lee (2010) "Program evaluation and research designs." In *Handbook of Labour Economics*, vol. 4a, pp. 463–536, eds. O. Ashenfelter and D. Card. New York: Elsevier
Djebbari, H., and J. Smith (2008) "Heterogeneous impacts in PROGRESA," *Journal of Econometrics* 145(1–2), 64–80
Dolton, P., and J. Smith (2011) "The econometric evaluation of the new deal for lone parents," IZA discussion paper no. 5491
Doolittle, F., and L. Traeger (1990) *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corp.
Doyle, J. J. (2008) "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care," *Journal of Political Economy* 116 (4), 746–70
Ford, R., D. Gyarmati, K. Foley, D. Tattrie, and L. Jimenez (2003) *Self-Sufficiency Project (SSP) – Can Work Incentives Pay for Themselves? Final Report on the Self-Sufficiency Project for Welfare Applicants*. Social Research and Demonstration Corporation (SRDC)
Forget, E. (2011) "The town with no poverty: The health effects of a Canadian guaranteed annual income field experiment," *Canadian Public Policy* 37(3), 283–305
Frölich, M., M. Huber, and M. Wiesenfarth (2015) "The finite sample performance of semi- and nonparametric estimators for treatment effects and policy evaluation," IZA discussion paper no. 8756
Goldberger, A. (1983) "Abnormal selection bias." In *Studies in Econometrics, Time Series, and Multivariate Statistics*, pp. 67–84, eds. S. Karlin, T. Amemiya, and L. A. Goodman. New York: Academic Press
Greenberg, D., and P. Robins (2008) "Incorporating nonmarket time into benefit–cost analyses of social programs: An application to the Self-Sufficiency Project," *Journal of Public Economics* 92, 766–94
Greenberg, D., and M. Shroder (2004) *Digest of the Social Experiments*. 3rd ed. Washington, DC: Urban Institute Press
Greenberg, D., M. Shroder, and M. Onstott (1999) "The social experiment market," *The Journal of Economic Perspectives* 13(3), 157–72
Guyatt, G. H., A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünemann (2008) "GRADE: An emerging consensus on rating quality of evidence and strength of recommendations," *British Medical Journal* 336(7650), 924–26
Heckman, J. (1979) "Sample selection bias as a specification error," *Econometrica* 47(1), 153–61
——— (1996a) "Randomization as an instrumental variable," *The Review of Economics and Statistics* 78, 336–41
——— (1996b) "Comment." In *Empirical Foundations of Household Taxation*, pp. 32–38, eds. M. Feldstein and J. Poterba. Chicago: University of Chicago Press
——— (2001) "Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture," *Journal of Political Economy* 109(4), 673–748
——— (2005) "The scientific model of causality," *Sociological Methodology* 35, 1–97
——— (2010) "Building bridges between structural and program evaluation approaches to evaluating policy," *Journal of Economic Literature* 48(2), 356–98
Heckman, J., P. Carneiro, and E. Vytlacil (2011) "Estimating marginal returns to education," *The American Economic Review* 101(6), 2754–871
Heckman, J., N. Hohmann, J. Smith, and M. Khoo (2000) "Substitution and dropout bias in social experiments: A study of an influential social experiment," *The Quarterly Journal of Economics* 115(2), 651–94

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998) "Characterizing selection bias using experimental data," *Econometrica* 66(5), 1017–98
Heckman, J., R. LaLonde, and J. Smith (1999) "The economics and econometrics of active labour market programs." In *Handbook of Labour Economics*, vol 3A, pp. 1865–2097, eds. O. Ashenfelter and D. Card. Amsterdam: North-Holland
Heckman, J., L. Lochner, and C. Taber (1998) "General-equilibrium treatment effects: A study of tuition policy," *The American Economic Review* 88(2), 381–86
Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010) "A new cost–benefit and rate of return analysis for the Perry preschool program: A summary." In *Cost-Effective Programs in Children's First Decade: A Human Capital Integration*, pp. 366–80, eds. A. Reynolds, A. Rolnick, M. Englund, and J. Temple. New York: Cambridge University Press
Heckman, J., and J. Smith (1999) "The pre-programme dip and the determinants of participation in a social programme: Implications for simple programme evaluation strategies," *The Economic Journal* 109(457), 313–48
——— (2000) "The sensitivity of experimental impact estimates: Evidence from the national JTPA study." In *Youth Employment and Joblessness in Advanced Countries*, pp. 331–56, eds. D. Blanchflower and R. Freeman. Chicago: University of Chicago Press for NBER
Heckman, J., J. Smith, and C. Taber (1998) "Accounting for dropouts in social experiments," *The Review of Economics and Statistics* 80(1), 1–14
Heckman, J., J. Tobias, and E. Vytlacil (2001) "Four parameters of interest in the evaluation of social programs," *Southern Economic Journal* 68(2), 210–23
Heckman, J., and S. Urzúa (2010) "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics* 156(1), 27–37
Heckman, J., and E. Vytlacil (2005) "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica* 73, 669–738
——— (2007a) "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation." In *Handbook of Econometrics*, vol. 6, pp. 4779–874, eds. J. Heckman and E. Leamer. Amsterdam: North-Holland
——— (2007b) "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 *Handbook of Econometrics*, vol. 6, pp. 4875–5143, eds. J. Heckman and E. Leamer. Amsterdam: North-Holland
Hertzman, C., A. Sweetman, R. Warburton, and W. Warburton (2014) "The impact of placing adolescent males into foster care on their education, income assistance and incarcerations," *Canadian Journal of Economics*, 47(1), 35–69
Hirano, K., G. Imbens, D. Rubin, and X. H. Zhou (2000) "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics* 1, 69–88
Ho, D., K. Imai, G. King, and E. Stuart (2007) "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political Analysis* 15, 199–236
Hotz, V. J., G. Imbens, and J. Mortimer (2005) "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics* 125(1–2), 241–70
Huber, M., M. Lechner, and C. Wunsch (2013) "The performance of estimators based on the propensity score," *Journal of Econometrics* 175, 1–21
Hum, D., and W. Simpson (1993) "Economic response to a guaranteed annual income: Experience from Canada and the United States," *Journal of Labour Economics* 11(1), S263–96

Ichino, A., F. Mealli, and T. Nannicini (2008) "From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity?," *Journal of Applied Econometrics* 23(3), 305–27
Imbens, G. (2010) "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature* 48(2), 399–423
——— (2015) "Matching methods in practice: Three examples," *The Journal of Human Resources* 50(2), 373–419
Imbens, G., and J. Angrist (1994) "Identification and estimation of local average treatment effects," *Econometrica* 62(4), 467–76
Imbens, G., and T. Lemieux (2008) "Regression discontinuity designs: A guide to practice," *Journal of Econometrics* 142(2), 615–35
Imbens, G., and D. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press
Imbens, G., and J. Wooldridge (2009) "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature* 47(1), 5–86
Kantarevic, J., and B. Kralj (2013) "Quality and quantity in primary care mixed payment models: Evidence from family health organizations in Ontario," *Canadian Journal of Economics* 46 (1), 208–38
Kline, P., and C. Walters (2015) "Evaluating public programs with close substitutes: The case of Head Start," NBER working paper no. 21658
Koremenos, B., and J. Smith (2015) "When to select a selection model in international relations research?," unpublished manuscript, University of Michigan
LaLonde, R. (1986) "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review* 76(4), 604–20
Lechner, M. (2010) "The estimation of causal effects by difference-in-difference methods," *Foundations and Trends in Econometrics* 4(3), 165–224
Lechner, M., and J. Smith (2007) "What is the value added by case workers?," *Labour Economics* 14(2), 135–51
Lee, D., and T. Lemieux (2010) "Regression discontinuity designs in economics," *Journal of Economic Literature* 48(2), 281–355
Leigh, A. (2009) "What evidence should social policymakers use?," *Economic Roundup* 1, 27–43
Lemieux, T., and K. Milligan (2008) "Incentive effects of social assistance: A regression discontinuity approach," *Journal of Econometrics* 142(2), 615–35
Levin, L., R. Goeree, N. Sikich, B. Jorgensen, M. Brouwers, T. Easty, and C. Zahn (2007) "Establishing a comprehensive continuum from an evidentiary base to policy development for health technologies: The Ontario experience," *International Journal of Technology Assessment in Health Care* 23(3), 299–309
Levitt, S., and J. List (2009) "Field experiments in economics: The past, the present, and the future," *European Economic Review* 53(1), 1–18
Lewbel, A., Y. Dong, and T. Yang (2012) "Comparing features of convenient estimators for binary choice models with endogenous regressors," *Canadian Journal of Economics* 45(3), 809–29
Lise, J., S. Seitz, and J. Smith (2004) "Equilibrium policy experiments and the evaluation of social programs," NBER working paper no. 10283
MacKinnon, J., and M. Webb (2016) "Randomization inference for difference-in-differences with few treated clusters," Queen's University Department of Economics working paper no. 1355
Manski, C. (2004) "Statistical treatment rules for heterogeneous populations," *Econometrica* 72(4), 1221–46
McCrary, J. (2008) "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics* 142(2), 698–714

Meyer, B. (1995) "Natural and quasi-experiments in economics," *Journal of Business & Economic Statistics* 13(2), 151–61
Milligan, K., and M. Stabile (2007) "The integration of child tax credits and welfare: Evidence from the Canadian National Child Benefit program," *Journal of Public Economics* 91(1–2), 305–26
Moffitt, R. (1991) "Program evaluation with nonexperimental data," *Evaluation Review* 15(3), 291–314
Morissette, R., X. Zhang, and M. Frenette (2007) "Earnings losses of displaced workers: Canadian evidence from a large administrative database on firm closures and mass layoffs," Analytical Studies Branch Research Paper Series, no. 291. Ottawa: Statistics Canada
Morris, P., and C. Michalopoulos (2003) "Findings from the Self-Sufficiency Project: Effects on children and adolescents of a program that increased employment and income," *Journal of Applied Developmental Psychology* 24(2), 201–39
Muller, S. (2015) "Causal interaction and external validity: Obstacles to the policy relevance of randomized evaluations," *The World Bank Economic Review* 29(Supplement 1), S217–25
Murray, M. (2006) "Avoiding invalid instruments and coping with weak instruments," *The Journal of Economic Perspectives* 20(4), 111–32
Neumark, D., and W. Wascher (2000) "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment," *The American Economic Review* 90(5), 1362–96
Oreopoulos, P. (2003) "The long-run consequences of living in a poor neighborhood," *The Quarterly Journal of Economics* 118(41), 533–75
——— (2006) "The compelling effects of compulsory schooling: Evidence from Canada," *Canadian Journal of Economics* 39(1), 22–52
Perez-Johnson, I., Q. Moore, and R. Santilano (2011) *Improving the Effectiveness of Individual Training Accounts: Long-Term Findings from an Experimental Evaluation of Three Service Delivery Models: Final Report*. Princeton, New Jersey: Mathematica Policy Research
Pitt, M., M. Rosenzweig, and N. Hassan (2010) "Human capital investment and the gender division of labor in a brawn-based economy," Yale Growth Center discussion paper no. 989
Riddell, C., and C. Riddell (2014) "The pitfalls of work requirements in welfare-to-work policies: Experimental evidence on human capital accumulation in the Self-Sufficiency Project," *Journal of Public Economics* 117, 39–49
Rothstein, J., and T. von Wachter (2015) "Social experiments in the labor market," unpublished manuscript, University of California at Berkeley
Roy, A. D. (1951) "Some thoughts on the distribution of earnings," *Oxford Economic Papers* 3, 135–46
Rubin, D. (2008) "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics* 2(3), 808–40
Schochet, P. (2008) *Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations. NCEE 2008–4026*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education
Sianesi, B. (2014) "Dealing with randomisation bias in a social experiment: The case of ERA," Institute for Fiscal Studies working paper no. W14/10
Smith, J., and A. Sweetman (2010) "Putting the evidence in evidence-based policy." In *Strengthening Evidence-based Policy in the Australian Federation*, Vol. 1: *Proceedings*, pp. 59–101. Canberra: Australian Productivity Commission

Smith, J., and P. Todd (2005) "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics* 125(1–2), 305–53
Sorkin, I. (2015) "Are there long-run effects of the minimum wage?," *Review of Economic Dynamics* 18, 306–33
Stock, J., J. Wright, and M. Yogo (2002) "A survey of weak instruments and weak identification in generalized methods of moments," *Journal of Business & Economic Statistics* 20(4), 518–29
Todd, P., and K. Wolpin (2006) "Assessing the impact of a school subsidy program in Mexico using a social experiment to validate a dynamic behavioral model of child schooling and fertility," *The American Economic Review* 96(5), 1384–417
van der Klaauw, W. (2008) "Regression-discontinuity analysis: A survey of recent developments in economics," *Labour* 22(2), 219–45
Warburton, W., and R. Warburton (2002) "Should the government sponsor training for the disadvantaged?" In *Towards Evidence-Based Policy for Canadian Education*, pp. 69–100, eds. P. de Broucker and A.Sweetman. McGill–Queen's University Press
Wing, C., and T. Cook (2013) "Strengthening the regression discontinuity design using additional design elements: A within-study comparison," *Journal of Policy Analysis and Management* 32(4), 853–77