4 Putting the evidence in evidence-based policy

Jeffrey Smith[1]
Department of Economics, University of Michigan

Arthur Sweetman[2]
School of Policy Studies, Queen's University

Abstract

Evidence-based policymaking presumes good evidence. This chapter considers what policymakers can do to enable and encourage the production of such evidence. The core of the chapter reviews the current state of knowledge on alternative ways of estimating the causal effects of programs and policies. While we highlight the value of social experiments, we also make clear that opportunities exist for increasing the quality of the evidence provided by non-experimental evaluations through improvements in policy design and implementation, in the collection of survey data and in administrative data systems. We lay out the role that policymakers can play in exploiting these opportunities. The chapter also considers programs that affect non-participants as well as participants, cost–benefit analysis, potential substitutes for econometric evaluation such as performance management and customer satisfaction, and institutional changes that could improve the quality of evaluation evidence.

4.1 Introduction

The success of evidence-based policymaking depends on the quality of the evidence that underlies it. In this chapter we consider how policymakers can improve the quality of the evidence they use in making policy decisions. We focus almost

[1] econjeff@umich.edu.

[2] sweetman@qsilver.queensu.ca.

exclusively on evidence regarding the effectiveness of programs at changing the outcomes of the individuals, firms or local governments they serve. Such evidence necessarily plays a key role in cost–benefit analyses designed to guide decisions about program initiation, expansion, contraction and termination. The different, but important, concerns addressed by audits and process evaluations lie outside our scope.

Our discussion emphasises that policymakers largely determine the quality of the evidence that they have available for making policy decisions. They exercise this control in a variety of ways, both direct and, more importantly, indirect. Good evidence depends on much more than just a demanding client, a topnotch evaluator and an adequate budget when commissioning an evaluation. It also depends on broader decisions about program design and implementation prior to evaluation, on the design and funding of general social science data sets, on the quality of administrative data systems, on peer review and on institutions that encourage the development of informed evaluation consumers within government. Policymakers must also avoid the temptation of thinking that performance management or customer satisfaction can substitute for rigorous impact evaluation.

The rapid pace of methodological development in program evaluation is the secondary theme of this chapter. Evaluations that looked good 15 years ago sometimes look mediocre now. Policymakers need to have some sense of the existence and nature of these developments. At the same time, we have endeavoured to keep the discussion accessible to relatively non-technical readers while providing numerous references to the literature for readers who want more depth.

We organise the chapter as follows:

- Section 4.2 considers parameters of interest in impact evaluation, building on the basic insight that programs and policies have effects that often differ substantially among those they serve, or over time or over space.
- Section 4.3 forms the main course of our intellectual meal. It reviews the basic experimental and non-experimental strategies for estimating the ‘partial equilibrium’ impacts of programs. We make the case for doing more experiments, but at the same time we highlight for each non-experimental methodology what policymakers can do to increase the quality of the evidence produced.
- Section 4.4 considers the difficult problem of accounting for spillovers and other ‘general equilibrium’ effects that arise when programs have effects on non-participants.

- Section 4.5 considers the notion of a 'hierarchy of evidence' that attempts to rank the different evaluation methodologies.
- Section 4.6 discusses ways to improve the practice of cost–benefit analysis while
- Section 4.7 critiques alternatives to impact evaluation that sometimes distract policymakers.
- Section 4.8 focuses on the role of data quality and what policymakers can do to improve it.
- Section 4.9 provides some suggestions for institutional improvements we think would increase evaluation quality.
- Section 4.10 concludes.

4.2 Parameters of interest

Policy discussions often casually refer to 'the effect' of a particular program as though it constitutes a universal constant. In fact, program effects often vary substantially along several important dimensions. The recent literature on evaluation has made remarkable progress both conceptually and empirically in clarifying the nature of heterogeneous program effects and tracing out the implications of heterogeneous effects for the design and execution of evaluations.

Before delving into the substantive issues raised by heterogeneous treatment effects, we need to lay down some conceptual and definitional foundations. We sometimes use 'units' as a generic term for participants to emphasise that programs may serve, say, firms or local governments, rather than individuals. We use the term 'treatment' as a generic term for programs and policies.

A treatment effect (sometimes called an 'impact' or just an 'effect') refers to the difference a treatment makes in the outcome of a unit. In this regard, we can think of each unit as having two outcomes: first, a treated outcome, realised in the (possibly counterfactual) world wherein the unit gets treated; and, second, an untreated outcome, realised in the (possibly counterfactual) world wherein the unit does not get treated. Some readers will recognise this as the so-called 'potential outcomes framework'. The treatment effect (sometimes called the 'causal effect') is the difference between these two potential outcomes. Put differently, the treatment effect consists of the value added to (or sometimes, subtracted from) the outcome as a result of treatment.

Most analyses focus on estimating averages of treatment effects across units. The average treatment effect on the treated (ATET), which provides an estimate of the expected difference between the treated outcome and the untreated outcome *for those who receive the treatment*, receives the most attention in the literature. This parameter informs a cost–benefit analysis that addresses the question of whether to keep or scrap a program in its present form. In contrast, the average treatment effect (ATE) estimates the expected effect of a program on all eligible units, whether or not they actually participate. This parameter informs a cost-benefit analysis that considers whether or not to make a program mandatory. For a program that is already mandatory, the ATET and the ATE coincide. For voluntary programs where potential participants have some idea of their own treatment effect, we expect positive selection on the treatment effect, so that ATET > ATE; put differently, we expect that in voluntary programs, participants will have higher impacts, on average, than all eligible units, while non-participants will have lower than average impacts. For voluntary programs, we might also be interested in impacts at the margin of participation; that is, impacts for those units for which a small change in costs or benefits would change the participation decision. The mean impact for these units (and not the ATET or ATE) should guide decisions about marginal expansions or contractions in the number of participants served. These impacts at the margin relate to what Imbens and Angrist (1994) named local average treatment effects (LATEs), which we discuss in Section 4.3.

In addition to varying by participation status, treatment effects may also vary among subgroups defined by observable characteristics, such as men and women, older or younger individuals, larger or smaller firms and so on. This variation may result from different patterns of selection across groups due to, for example, differences in the cost of participation. If two groups have the same distribution of treatment effects, but different costs of participation, the group with a lower cost of participation should have a higher participation rate but a lower average treatment effect conditional on participation. Variation across groups may also result from differences in the appropriateness of the treatment, what we might call the match between the treatment and the group. For example, a textbook-based job search course may have a larger treatment effect on more educated participants due to their presumably greater facility at absorbing written material.

In some evaluation contexts, such as educational interventions and active labour market programs, presentation of subgroup impacts has become fairly standard. Differences in impacts across groups that result from differences in the quality of the match between the treatment and the group may illuminate aspects of program operation not obvious from the overall impacts and so suggest where to focus efforts at program reform. Such differences in impacts can also guide efforts to use statistical treatment rules to target program services, as described by Lechner and Smith (2007) and Blattman (2008), or as in the survey by Smith and Staghøej (2010). Such rules formalise the assignment of treatments based on characteristics that predict larger treatment effects. Where sample sizes allow it, policymakers should want to see, and evaluators should want to provide, subgroup impacts.

Average treatment effects may also vary in other ways. For example, the impacts of active labour market programs may vary over the business cycle, as described by Lechner and Wunsch (2009). They may remain stable over time after participation as with the US Job Training Partnership Act (US General Accounting Office 1996) or they may fade out over time as with the California Greater Avenues to Independence (GAIN) program (Hotz, Imbens and Klerman 2006). Impacts may vary by local or regional office due to local social or economic conditions or due to variation in the quality of program management or program staff. These types of variation should interest policymakers as well; they imply a concern at the evaluation design stage with ensuring the availability of adequate sample sizes for precise estimation of differential impacts over time or across offices or regions. Finally, treatment effects may vary across units even within subgroups, or time periods, or local offices; Heckman, Smith and Clements (1997), Bitler, Gelbach and Hoynes (2006) and Djebbari and Smith (2008), among others, address the related conceptual and econometric issues.

To summarise, the fact that treatment effects vary across units has important implications for evaluation design, execution and interpretation. Different mean treatment effects address different policy questions and often suggest different econometric evaluation strategies. Careful evaluation design should lead to harmony among these elements. In addition, designing in sufficient sample size to capture variation in treatment effects across key dimensions often adds great value to the findings from evaluations.

4.3 Partial equilibrium evaluation methods

This section lays out the standard econometric methods used to estimate the impact of interventions in a ‘partial equilibrium’ context. Partial equilibrium is economist-speak for operating under the assumption that the program only affects participants, and so does not affect non-participants via spillovers such as displacement in the job market or changes in market prices. Section 4.5 considers such spillover effects.

Social experiments

Random assignment and the selection bias problem

To see the problem that random assignment solves, think about estimating the ATET for some voluntary program. The ATET consists of the difference between the observed mean outcome of program participants and the counterfactual (and thus unobserved) mean outcome that program participants would have experienced had they not participated. The first of these presents little trouble to the evaluator, as it requires only collection of outcome data on a random sample of participants. The second of these presents the evaluator with a much more difficult problem: how to estimate what would have happened to participants in the imaginary world in which they did not actually participate. We cannot simply draw a random sample of eligible non-participants and estimate their mean outcome because we worry (quite rightly and with much evidence to back up our concerns) that individuals select non-randomly into programs. As a result, a comparison of the outcomes of participants with the outcomes of eligible non-participants will conflate the effects of the program (if any) with other differences between participants and non-participants that would have shown up in outcomes even if the program did not exist and the participants experienced their non-participation outcomes. For example, the participants might have higher levels of education, ability or motivation, or be better looking, or just have fewer other things, like young children, to keep them busy and so away from the program. The literature calls bias that results from non-random program participation 'selection bias'.

Randomisation solves the selection bias problem by taking a group of would-be program participants and randomly forcing some of them to realise their untreated outcome by excluding them from the treatment. In samples of reasonable size, the units randomly assigned to the treatment group and allowed to receive treatment will have the same pre-program characteristics, both observed and unobserved, as the randomly assigned control group that gets excluded from treatment. As a result, the mean difference in outcomes between the experimental treatment and control groups provides an unbiased estimate of the ATET. Random assignment makes the two groups statistically equivalent in all aspects other than access to treatment, with the result that only the difference in treatment can cause a difference in outcomes between them.

In the United States, 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, the negative income tax, and abstinence-only sex education. Greenberg and Schroder (2004) document all but the most recent US experiments. The last few years have also witnessed an explosion in experiments in developing countries (see for example, Banerjee and Duflo 2009). As a result of all these experiments, a large body of knowledge regarding design and implementation exists as well as many organisations capable of pulling off high-quality experimental evaluations. Thus, policymakers in countries with little experience with experimental evaluation have little to fear and much to gain.

Issues with random assignment

Burt Barnow likes to say that random assignment is not a substitute for thinking; on this theme see the article by Barnow (2010) and also the humorous but pointed contribution by Smith and Pell (2003). Indeed, experiments present a more difficult evaluation challenge than their basic conceptual simplicity suggests. Experiments accomplish one very important thing: they solve the selection bias problem in partial equilibrium evaluations in a simple and compelling way. As noted by Heckman and Smith (2000), experiments remain subject to all the other issues that make empirical program evaluation so much fun, such as outliers, survey non-response and attrition, error-filled and poorly documented administrative data, and Hawthorne effects. Experimental evaluations may also have issues with external validity, particularly when relying on volunteer sites. 'External validity' refers to the extent to which program impact estimates obtained at a given time and in a particular place plausibly carry over to other times and places (while 'internal validity' refers to the applicability of the estimates to the time and place of the evaluation).

As discussed in detail in section 5 of Heckman, LaLonde and Smith's chapter (1999), experimental evaluations also face some issues typically not faced in non-experimental evaluations. Non-experimental evaluations compare treated units to untreated units using the various methods discussed under 'Selection on observed variables' and 'Regression discontinuity' in this section. In contrast, experiments often randomly assign *access* to treatment, rather than treatment itself, with the result that in many contexts, not all experimental treatment group members actually receive treatment and, less often, some control group members obtain the same or similar treatments from other sources. In the presence of treatment group dropout, the difference in outcomes between the treatment and control groups now estimates the mean impact of the offer of treatment (called in the literature the 'intention to treat') rather than the mean impact of treatment itself. Things get even more complicated with control group substitution into similar services from other sources. The articles by Heckman, Smith and Taber (1998) and Heckman, Hohmann, Smith and Khoo (2000) and the 'Instrumental variables' section below discuss these issues in greater depth.

An experimental evaluation may require a program to dig deeper into its eligible population than it normally would in order to fill up the control group while still maintaining its normal scale of operations. In such cases, external validity concerns arise because the participant population during the experiment differs from the usual participant population. Also, randomised rather than deterministic access to treatment may deter complementary investments prior to treatment or may change the composition of participants by deterring the risk-averse and attracting the risk-loving, again raising issues of external validity.

Though very real and of serious concern, thoughtful experimental design can often reduce the practical importance of these concerns; only occasionally do they become severe enough to outweigh the general case for random assignment.

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 a program. Such uses address different questions that sometimes possess equal or greater policy relevance and often avoid or reduce political, practical and ethical (see below) concerns related to a no-treatment control group. Consider two illustrative real world examples.

Black, Smith, Berger and Noel (2003) document the clever use of randomisation in the Unemployment Insurance (UI) system in Kentucky. Like all other US states, Kentucky employs a statistical model to predict the fraction of the (usually) 26 weeks of UI benefit entitlement each new claimant will consume as a function of claimant and local area characteristics. The state then converts this predicted duration into a score between one and twenty, with twenty indicating benefit eligibility exhaustion and one indicating a very short predicted duration. In each local UI office in each week, the state assigns new UI claimants to receive (or not) mandatory reemployment services based on their score. Assignment starts with the highest score in a given office and a given week and proceeds until it runs out of slots or claimants. In many 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; these slots are randomly assigned. This scheme passed the scrutiny of sceptical state officials who were concerned that the alternative of randomly assigning all claimants, including those with long predicted durations on UI, would break the state budget.

The ‘randomisation 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 just 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 current policy interest.

McConnell, Decker and Perez-Johnson (2006) experimentally evaluate three alternative ways of structuring the 'Individual Training Accounts' (ITAs) provided to some participants in the US Workforce Investment Act (WIA) program. The three alternatives 'vary in whether counseling is mandatory, whether the counselor is asked to direct the participant in their training choice and can veto the participant's ultimate choice, and whether the value of each ITA is preset or determined by the counselor.' Everyone receives services but important aspects of the service delivery process differ among the three 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. Other variants of random assignment include randomised rollout of programs too big to put in place in all locations at the same time, and randomised encouragement designs, as described by Hirano, Imbens, Rubin and Zhou (2000), that randomly assign not treatment but an incentive to participate in the treatment.

In short, given the tremendous variety of possible randomised designs, we can hardly overemphasise the potential to conduct persuasive yet inexpensive (and relatively uncontroversial) experimental evaluation.

Ethics, politics and experiments

Policymakers sometimes express ethical concerns with the random service denial inherent in random assignment designs with 'no treatment' or even 'less treatment' control groups. In our view, these ethical concerns often simply provide cover for policymakers who prefer not to have clear evidence on program effectiveness, perhaps because they think the program would not pass a benefit–cost test even though it succeeds in transferring public resources to favoured groups such as providers or clients.

While noting the potential for ethical misrepresentation, advocates of experiments can also 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 having its time and effort wasted on an ineffective treatment. Second, the government can always compensate experimental participants for contributing to the public good of knowledge creation. Unlike the case of some medical treatments, only modest payments should quell any ethical concerns for most social policies. Third, an alternative and perhaps weightier ethical concern militates in favour of random assignment where possible. Is it really ethical for policymakers to spend public money (implicitly taken by force from taxpayers) on programs without a compelling evidentiary basis, when they could easily bring about the production of such evidence?

Selection on observed variables

Consider the case where non-random selection into treatment occurs but the analyst observes all the variables with important effects on both participation and on the outcome of interest in the absence of participation. Economists call this case 'selection on observed variables' while statisticians call it 'unconfoundedness'.

Selection on observed variables represents a very strong assumption indeed! In our view, most evaluations that rely on this assumption fall far short of this standard, sometimes because of data limitations and sometimes, more broadly, because we simply lack the knowledge in many policy contexts of what variables to condition on. Successful application of this strategy requires careful thought about the institutions and the economics of the situation in order to make the case that all of the variables that both theory and existing empirical knowledge suggest should appear among the conditioning variables in fact do so. Making this case requires much more than just saying that the evaluation uses 'rich' data containing a large number of variables, though many evaluations offer up only this unconvincing justification. It is not the number of conditioning variables that matters, but rather having the ones that make the 'selection on observed variables' assumption plausible.

When relying on the selection on observed variables assumption, analysts typically employ either a parametric linear regression model or else some sort of weighting or matching estimator, such as inverse probability weighting or propensity score matching. In general, weighting and matching estimators represent the first choice for various technical reasons, provided the sample size justifies their use. See, for example, the methodological discussions by Heckman, Ichimura, Smith and Todd (1998), Angrist (1998), Smith and Todd (2005), Caliendo and Kopeinig (2008) and Busso, DiNardo and McCrary (2009a, 2009b).

Policymakers and evaluators can take many steps to make the evidence provided by evaluations based on the selection on observed variables assumption 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 matching variables. Process evaluations can provide further information about the factors influencing participation decisions. Collecting data on factors that often go unmeasured, such as the attitudes toward work, future orientation (that is, discount rate), risk aversion, motivation, social and other non-cognitive skills, and cognitive ability of potential program participants, could also make the selection on observed variables assumption more credible.

A larger literature suggests the value in many substantive contexts of flexibly conditioning on past outcomes measured at a relatively fine level of temporal detail. In the context of active labour market programs, see, for example, the articles by Card and Sullivan (1988), Heckman, Ichimura, Smith and Todd (1998) and Dolton and Smith (2010). Collecting such data, or obtaining it from administrative records, is often a more or less necessary condition for relying on methods that assume selection on observed variables. In addition, policymakers can require formal sensitivity analyses along the lines of those in the articles by Altonji, Elder and Taber (2005) and Ichino, Mealli and Nannicini (2008) that indicate the inferential consequences of departures from the selection on observed variables assumption. Finally, and perhaps most importantly, policymakers can fund basic social science research on the determinants of participation and outcomes that provide the foundation for choices about data collection and analysis and for arguments about the plausibility of the selection on observed variables assumption for particular combinations of treatment, data and outcomes.

Instrumental variables

Instrumental variables (IV) can sometimes provide consistent estimates in contexts where selection into a program occurs on variables unobserved by the analyst, rendering the methods described above under ‘Selection on observed variables’ inappropriate. An ‘instrument’ (nothing to do with marching bands) is a variable that affects participation in the program but is not correlated with outcomes other than through its affect on participation. The classical bivariate normal selection model for which Heckman (1979) developed a famous estimator represents a close cousin to IV; all of the same comments apply.

A good instrument has two properties. First, it strongly predicts treatment receipt, where the recent technical literature precisely defines how strong is strong enough. On this point, see the oft-cited paper by Bound, Jaeger and Baker (1995) and the literature it spawned. This property has the pleasant feature that it lends itself to easy testing using the available data. The better the instrument predicts participation, the more powerful (in the statistical sense) the analysis for a given sample size or, put the other way, the stronger the instrument the smaller the sample required to obtain a given level of statistical power.

Second, a valid instrument affects outcomes only through its effects on the treatment, conditional on the included covariates. For example, intellectual ability does not represent a good instrument for schooling in an analysis of labour market outcomes, because while intellectual ability has a strong positive relationship with schooling attainment, and so possesses the first property of a good instrument, it also affects labour market outcomes directly, conditional on years of schooling. Put differently, even within groups with the same amount of schooling, intellectual ability will still predict labour market outcomes, and so it lacks the second property of a valid instrument. In contrast, random assignment yields an ideal instrument in the form of the indicator variable for belonging to the treatment group. By construction, this variable predicts treatment but has no relationship to outcomes other than through its effect on treatment. The search for instrumental variables in policy evaluation represents a search for variables that embody similarly random variation in program participation.

In general, there is no way to test the second property of a good instrument short of running an experiment. Instead, the analyst must 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. This process of argumentation renders instrumental variable estimates controversial in many contexts. Chapter 4 of Angrist and Pischke’s book (2009) provides a good conceptual introduction to instrumental variables. Blundell, Dearden and Sianesi (2005) explicate and apply IV methods (as well as the bivariate normal selection model) in the context of a study of the effects of schooling on labour market outcomes. Heckman, Tobias and Vytlacil (2001) provide a broad conceptual framework for thinking about instruments.

Where do good instruments come from? Sometimes nature provides instruments, as when Kochar (1999) uses weather as an instrument for agricultural income or when the sex composition of the first two children serves as an instrument for the total number of children in the Angrist and Evans study (1998) of the effect of children on women’s labour supply. Sometimes social events provide an instrument as in the Evans and Lein study (2005) that uses a bus strike in Philadelphia to study the impact of prenatal care on low income mothers. In other contexts, nature and institutions combine as in the paper by Evans and Kim (2006) that uses random variations in emergency room admissions on the weekend to study the impact of nurse-to-patient ratios on patient outcomes. And sometimes government itself provides an instrument, as with the variation in funding levels between jurisdictions that cut across the same local labour market employed in the study by Frölich and

Lechner (2010). In each of these cases, the researchers can make a good case that their instrument has both of the properties of a good instrument described above.

The literature on applied econometrics has spent the last decade or so coming to grips with the fact that analyses using instrumental variables generally estimate a somewhat unusual treatment effect parameter. In particular, under some (usually innocuous) assumptions they estimate the impact of the treatment on those whose treatment choice depends on the value of the instrument. Economists call this the 'local average treatment effect' (LATE) and statisticians call it the 'complier average causal effect' (CACE) where the compliers are those who change their treatment status when the instrument changes.

It helps to consider a couple of examples. In an experiment with treatment group dropout and control group substitution, the LATE is the impact on those who would receive treatment if assigned to the treatment group but not if assigned to the control group. In the context of the bus strike paper cited above, the analysis estimates the mean impact of prenatal care on those who would obtain prenatal care when there is not a bus strike, but who do not obtain it when there is a bus strike. Or consider the literature that uses variation over time or over jurisdictions in the compulsory schooling age to estimate the labour market impact of additional schooling, such as the article by Oreopoulos (2006). 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 16 to 17 years in the North American 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 go to college or university regardless of the value of the compulsory schooling age.

It follows quickly from the insight that each instrumental variable estimates a LATE to the insight that different instrumental variables will estimate LATEs corresponding to different complier groups. Some instruments will estimate LATEs of great relevance to policy, while others will not. In general, no instrument will estimate the ATET parameter, which means that instrumental variables estimates typically cannot directly answer the 'keep it or cut it' question that underlies most 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 relevant policy dimension consists of modest spending increases to reduce the costs of program access (or small cuts that would increase those costs). Recent papers by Deaton (2009), Heckman and Urzua (2009) and Imbens (2009) debate this and related issues.

The quality of the estimates obtained by applying IV methods depends on the quality of the instrument. A weak or invalid instrument may be worse than no instrument. Good instruments can be obtained in one of three ways: clever data collection, exploitation or creation of useful institutional variation, and randomisation. Obtaining good instruments is facilitated by careful planning at the program design and implementation stage (to produce that useful institutional variation) and at the time of evaluation design. Collection of high-quality data aids in instrumental variables analyses as well, whether because the data contains potential instruments or because having better conditioning variables available makes it more plausible to assume that an instrument generated outside the data (that is, from institutional variation) satisfies the second property discussed above.

Longitudinal methods

Longitudinal methods use variation over time in treatment status to estimate the impact of treatment. The simplest longitudinal method consists of a comparison of outcomes before and after treatment. This before–after estimator can be applied to individuals, as when comparing outcomes before and after participation in a training program, or to a jurisdiction, as when comparing alcohol-related fatalities at the state level before and after a change in the minimum legal drinking age. The implicit assumption underlying before–after comparisons is that in the absence of the treatment or policy change, outcomes in the 'after' period would have been the same as (at least in expected value terms) the 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. For example, in the training program case, an individual might choose to participate in training following job loss. If the individual would have found a job reasonably quickly even without training, then a before–after comparison that includes the period of unemployment prior to the start of training produces an upwardly biased estimate of the effect of training on earnings. In the case of the minimum legal drinking age, a change in the fraction of the population between the ages of 18 and 22 or changes in related policies, such as the blood alcohol level used to define drink-driving, at around the same time might confound a causal interpretation of the before–after outcome difference.

Concerns about the plausibility of simple before–after comparisons have led many researchers to prefer the 'difference-in-differences' estimator. 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. This estimator 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 common trends in outcomes. Both difference-in-differences and more general panel data studies rely on the assumption that, in the absence of the program or policy, the beforeafter change in outcomes for the treated units would equal (at least in expectation) that for the untreated units. Put differently, any differences between the treated units and the untreated units must remain constant between the before and after periods or, in the case of more general panel models, over the period covered by the data. Some parts of the literature refer to this situation (perhaps a bit misleadingly) as a 'natural experiment'; for further discussion see for example, Meyer's article (1995).

In certain contexts, the assumption of a common change in expected outcomes between treated and untreated units in the absence of treatment will make sense when an assumption of no change in outcomes in the absence of treatment for the treated units would not. At the same time, difference-in-differences is not a panacea. In cases where the treated units select into treatment based on transitory outcome changes, the difference-in-differences assumption fails. Thus, much of the intellectual action when considering evaluating a program or policy using these methods centres on learning about how the treated units came to be treated *when* they did. The analyst must also worry about anticipatory effects in the form of 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 of studies from the literature that use this method will help to clarify the picture, and to illustrate the many different types of comparison groups employed within this estimation framework. Heckman and Smith (1999) apply difference-in-differences in the context of a job training program. The comparison group consists of eligible non-participants in the sample local labour markets as the participants. Using an experimental benchmark, they find that that difference-in-differences performs poorly in this context, exhibiting both bias and strong sensitivity to the choice of particular 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) provides an example of difference-in-differences 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 evaluation (2007) of changes to Canada's National Child Benefit using both differences-in-differences across provinces provides another example using jurisdictional policy variation.

In the United States, state level policies ranging from right-to-carry (a gun) laws to minimum legal drinking ages have had their effects estimated via panel data models applied to state level data on policies and outcomes. Many of these studies fail to do much to justify the application of these methods, which is to say that they do little to convince the reader that the timing of policy changes at the state level does not depend on transitory changes in the outcomes of interest. The United States has something of an advantage in the application of panel data methods to policy evaluation compared to countries with smaller numbers of jurisdictions because it has 50 states rather than six states as in Australia or 10 provinces as in Canada. This additional variation provides useful degrees of freedom and leads directly to a recommendation to the governments of countries like Australia and Canada to break up large states and provinces into smaller ones so as to facilitate policy experimentation and evaluation (!).

Discussions of the econometrics of longitudinal evaluation methods can be found throughout the literature. Moffitt (1991) provides an accessible introduction. Wooldridge (2002), Cameron and Trivedi (2005) and Angrist and Pischke (2009) provide textbook treatments. Bertrand, Duflo and Mullainathan (2004) highlight important issues regarding calculation of the standard errors when applying longitudinal methods. Heckman and Hotz (1989) highlight the use of additional periods of data to do tests of the assumptions underlying longitudinal evaluation methods. Heckman (1996) critiques the application of difference-in-differences methods.

Policymakers have almost complete control over the ability of researchers to apply longitudinal methods to the evaluation of treatments at the jurisdictional level. Many programs roll out over time rather than all at once to avoid administrative overload and to allow later implementing jurisdictions to learn from the early movers. Randomly assigning the order in which jurisdictions implement a program, as in the rollout of the PROGRESA conditional cash transfer program in Mexico, represents a gold standard. Absent random assignment, trying to avoid rolling out programs in a way that is correlated with the outcomes it is designed to affect is a second best. Regardless of what is done, carefully documenting the decision rule used to order the rollout and the actual timing of implementation on the ground at least gives the evaluator a fighting chance.

Beyond program implementation, the ongoing collection of large social science panel datasets that include information on the geographic location of respondents, along with detailed information on program participation and outcomes at a relatively fine level of temporal detail, facilitates the application of longitudinal methods to the evaluation of both individual level treatments and jurisdiction level treatments. Panel data sets represent a complement to, rather than a substitute for, quality administrative outcome data at the jurisdictional level, as with data on alcohol-related traffic deaths or receipt of income assistance. A key in both cases is having data on outcomes that begins prior to the treatment under study.

Regression discontinuity

Regression discontinuity (RD) designs exploit discontinuous changes in treatment receipt that result from discontinuities in program rules. The RD estimator 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 value of the continuous variable, a comparison of mean outcomes just above and just below the cut-off value can provide a compelling source of information about treatment effects. The literature calls the continuous variable that determines treatment assignment the 'running variable' and the particular point at which the probability of treatment changes abruptly the cut-off value or discontinuity (from which comes the name regression discontinuity). The econometric literature defines a number of different estimators for the RD case, but they all just represent different ways of taking 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 what the literature calls 'sharp' and 'fuzzy' RD designs. In a sharp design, the probability of treatment moves from zero to one the discontinuity point. 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 vary discontinuously at the cut-off. For example, a senior citizen discount on publicly provided flu shots could induce a discontinuity in the probability of receiving a flu shot at age 65. In the United States, the distribution of ages at which children start the first year of primary school corresponds to a fuzzy design at the nominal age cut-off due to selective choices by parents and administrators to advance or delay particular children relative to the norm. In the fuzzy case, under certain pesky but often plausible additional assumptions, one can estimate the LATE on those units who change their treatment status at the cut-off value. 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 without the discount. It does not provide information about the impact of a shot on those who would get one with or without the discount, or who would not get one with or without the discount.

In both the sharp and fuzzy cases, generalisation of the estimated impacts to units with values of the running variable other than the value at the cut-off requires additional assumptions; the plausibility of such assumptions will depend on both prior knowledge, such as how the mean outcome varied with the running variable in periods prior to the implementation of the treatment, and on the institutional context.

A few examples will help to clarify the mechanics and the usefulness of RD. Perhaps the most well-known RD evaluation in the US context is that of the Reading First program commissioned by the Institute for Education Sciences of the US Department of Education and executed by Abt Associates and MDRC; see the final report by Jackson et al. (2007) for more information. This evaluation relied on the discontinuity created by the use of an index score to assign Reading First grants and found no real effect of Reading First on the outcomes of primary school children. Of course, these results apply only to schools near the discontinuity, a point missed in much of the discussion surrounding the evaluation, including the Dillon article (2008) in the *New York Times*.

Lee and McCrary (2009) examine the effect of punishment severity on criminal behaviour using the discontinuity in the US legal system between the punishment regime for juveniles (age less than 18 years) and adults (age 18 and above). Their setup has two main virtues: lots of data around the discontinuity and a very strong treatment due to large differences in severity between the juvenile and adult punishment regimes. Indeed, much to the surprise of pretty much everyone, they find only a very small change in criminal behaviour at age 18, indicating, at least for this age group, either a very present-oriented outlook or a very small response to anticipated punishment or both.

The foundational papers in economics are by Goldberger (1972a, 1972b, 2008), and consider a compensatory education program allocated according to a test score, with students scoring below a cut-off assigned to the program and those scoring above the cut-off not. Cook (2008) gives a broad history of RD in the social sciences. For methodological details on RD see the surveys by van der Klaauw (2008), Imbens and Lemieux (2008), and Lee and Lemieux (2009) and chapter 6 of Angrist and Pischke's book (2009). McCrary (2008) provides a useful test of the assumption of no manipulation of the running variable around the cut-off.

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.

Successful use of a discontinuity design in program evaluation demands more than just a discontinuity in program eligibility rules or in the costs of program participation. The discontinuity must build on a 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 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.

When relying on an age cut-off in a discontinuity analysis, the analyst must address the potential for spillovers, as in De Giorgi's study (2008) of the British New Deal for Young People (NDYP). His analysis relies (in part) on comparing the labour market outcomes of young unemployed people just above and just below the age cut-off for NDLP eligibility. To the extent that these young people represent close substitutes in the labour market, we would expect the existence of the program to have effects on both. Using either calendar time or age to define a discontinuity also raises the potential for anticipatory behaviour that has the potential to bias the estimated treatment effects.

Program designers need to locate the discontinuity at a point with many potential participants, so that sufficient data will exist to estimate a treatment effect with reasonable power; a sometimes difficult standard to reach given that sample sizes required for discontinuity designs typically exceed those for randomised trials with comparable power, as documented by Schochet (2008). Finally, the discontinuity in the policy variable must generate a corresponding discontinuity in treatment receipt. These criteria represent a tall order for program designers, and even when satisfied the evaluation still yields (as noted earlier) an estimated treatment effect only at the discontinuity. At the same time, a well-executed evaluation using regression discontinuity methods has nearly the same credibility as a well-executed experiment.

Other partial equilibrium evaluation methods

It is worth briefly commenting on some other partial equilibrium evaluation approaches. Process evaluations, such as the fine examples by Doolittle and Traeger (1990) and Kemple, Doolittle and Wallace (1993) from the US Job Training Partnership Act (JTPA) experiment, examine the flow of money and participants within programs. They have great value, but represent a complement to, rather than a substitute for, the sort of impact evaluation considered in this chapter. We have much the same view of comparative case studies, which can add richness to our understanding of outcome differences between programs or between sites within a program but cannot substitute for large sample econometric evaluations. Lab experiments have also started to play a small role in the evaluation literature (see, for example, Eckel, Johnson and Montmarquette 2005, Eckel et al. 2007, and Falk and Fehr 2003). In our view, lab experiments have the potential to play a small but useful role in evaluation going forward, though it will take some time for the lab experimenters to learn to think like evaluators and for evaluators to learn that lab experiments present more challenges than it might appear from outside. The hierarchical linear models, sometimes called multilevel models, widely used in education research (see, for example, Raudenbusch and Bryk 2001) represent not a separate method but rather a particular framework within which to apply the methods already described. This approach has the advantages of focusing attention on correct calculation of the standard errors for group (usually classroom or school) level treatments, of encouraging careful thought about causal relationships across institutional levels and of highlighting heterogeneity in the effects of treatments. Finally, structural methods (as economists use that term) rely on economic theory and related functional form assumptions to fill in for missing data. In the right hands the structural approach can add powerfully to the approaches already described. Todd and Wolpin (2005) provide an excellent example of the partial equilibrium structural approach.

4.4 Spillovers and general equilibrium effects

We now consider the effects of programs on persons or organisations or markets that do not directly participate in them. Such spillovers may accrue directly, as when a training program improves life for the family members of the trainee or an educational intervention reduces crime, or indirectly, via the operation of labour and product markets (or even changes in norms). Economists refer to indirect spillovers as general equilibrium effects. For example, a job placement program that helps one group of people find jobs may simultaneously make job finding more difficult for another group for whom they represent substitutes in production. Ethanol subsidies in the developed world may drive up the price of food in developing countries.

These external costs and benefits 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 on achievement in the subjects whose classroom time gets reduced. Evaluations of labour market programs should collect data on criminal behaviour, as in the US National Job Corps Study, where reductions in crime represent an important component of program impacts, and on children, as in the Morris and Michalopoulos analysis (2003) of Canada’s Self-Sufficiency Project.

Evaluators can sometimes obtain estimates of indirect spillovers by assigning treatment at the group level and then measuring outcomes for both participants and non-participants. For example, Dahlberg and Forslund (2005) use variation across municipalities to estimate the displacement effects of wage subsidies (large) and training (small) in Sweden. Many educational interventions affect some but not all students in a classroom; assigning the intervention to classrooms rather than students and then measuring outcomes of all students in both treated and untreated classrooms allows estimation of any spillovers. Finally, the clever village-level random assignment in the experimental 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.

In many cases, obtaining estimates of general equilibrium effects will require writing down and either estimating or calibrating a 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 evaluation business. This means it does not make sense to undertake such ventures for every evaluation of every program. Instead, general equilibrium evaluation analyses of this type should address important cases in terms of program size or program design and proceed on a somewhat separate track (that is, with more attention from academic economists and more funding from research funders rather than policy funders). Most evaluations should simply draw on this broader literature when discussing the possible nature and extent of such effects in a given context and when examining the sensitivity of cost–benefit calculations to likely general equilibrium effects.

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) looked for displacement effects in one of the US Unemployment Insurance (UI) bonus experiments. In these experiments, the treatment consisted of the offer of a cash bonus to claimants who found a job early in their UI spells. They estimate that the displacement of workers not in the experiment cancelled out about 20 per cent of the employment impact of the program estimated in the experiment. In a study of tuition subsidy programs for university students, Heckman, Lochner and Taber (1998) find much larger general equilibrium effects. In their study, the partial equilibrium estimate of the impact of treatment on the treated is ten times larger than a general equilibrium impact that accounts for the decline in the relative wage of persons with a university degree resulting from their increased supply. Finally, Lise, Seitz and Smith (2010) examine the general equilibrium effects of Canada's Self-Sufficiency Project, an earnings subsidy for single parents on income assistance for at least a year who find a job during the second year of their benefit receipt spell. 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) in the labour market leads to a reversal of the positive cost–benefit conclusions reached in the partial equilibrium experimental evaluation.

Policymakers play a limited but important role here. Discussion of possible direct and indirect spillovers should take place when designing an evaluation's basic identification strategy and when laying out plans for data collection. Policymakers should insist on such discussions at the start of each evaluation and should make sure that spillovers play a role in the interpretation of the impact estimates and in the related cost–benefit calculations at the end of the evaluation as well.

4.5 Comparing and ranking econometric evaluation methods

Leigh (2009) draws on a literature that proposes various 'hierarchies of evidence' and proposes his own hierarchy for Australia (see his Box 3). A generic version of such a hierarchy would have random assignment studies on top, followed by regression discontinuity designs, followed by instrumental variables or difference-in-differences designs, followed by studies relying on selection on observed variables, followed by before–after comparisons, expert opinion and, at the very bottom (what would our theorist colleagues say?), 'theoretical conjecture'.

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 published evaluation studies using each method would likely correspond to this ordering. Nor do we dispute that this information has some value. Our concern lies in two not uncommon misinterpretations of such rankings. First, this ranking focuses on the 'between' variation rather than the 'within' variation, which leads some observers to forget the 'within' variation entirely. In fact, the relative importance of differences in the average quality of evaluations using the various different methods and variation in quality conditional on method is an empirical question, one well worth investigating and one for which we know of no available systematic quantitative evidence.

Second, the differences in mean quality across methods represent an equilibrium relationship; they need not be causal in the sense that, in a given context, moving up the hierarchy may make things worse rather than better. A given study, for example, may rely on cross-sectional data and an assumption of selection on observed variables because, in its context, no good instrumental variables suggest themselves and, looking at the time series of outcomes, there appears to be important selection into treatment based on transitory shocks. In this case, moving 'up' the hierarchy will likely lower the quality of the evaluation because it will mean using an invalid instrument or applying longitudinal methods when the assumptions that underlie them fail to hold in the data.

This second concern leads us directly to the misguided literature set in motion by LaLonde (1986). This literature seeks the holy grail of non-experimental evaluation: a non-experimental method that always and everywhere solves the selection problem. Dehejia and Wahba's works (1999, 2002) represent the most famous papers in this literature, which many (not necessarily including the authors) have interpreted as showing that matching 'works' in the sense of always solving the selection problem. Their work in turned spawned a large literature addressing the question of 'does matching work?' by comparing matching estimates to experimental estimates, sometimes using laughably weak sets of conditioning variables in the matching. In fact, the question 'does matching works' is ill posed. As noted in Section 4.3 under 'Selection on observed variables', matching works in the sense of providing consistent estimates when the available variables suffice to make the conditional independence assumption hold in a given context and not otherwise. Thus, we know the answer to the generic 'does matching works' question in advance; it is 'sometimes, but only when the data support it.'

Put in the context of our discussion of hierarchies, sometimes matching will outperform methods ranked above it in the hierarchy of evidence, as in a context where the analyst observes all the relevant conditioning variables but no instruments, and sometimes not. Rather than searching for a non-existent magic bullet estimator the literature should seek to build a body of knowledge on what methods work for particular combinations of parameter of interest, available data, and program institutions. Rather than relying on a hierarchy to choose an identification strategy, researchers should seek to use the particular strategy best suited to providing a compelling impact in a given context given the nature of the program institutions and the data at hand. In our view, the main role of evidentiary hierarchies is to give policymakers an extra nudge in favour of experiments and to encourage them to push hard on evaluators who claim that a strategy low on the hierarchy represents the best choice in a given context.

Two final points on hierarchies: First, one should, of course, use all of the available high-quality evidence rather than just relying on one study. Meta-analysis represents a very useful tool for combining evidence, but it does not create any new evidence. Thus, it seems out of place in Leigh's (2009) hierarchy of evidence for Australia. Moreover, as poorly done meta-analyses often obscure the high-quality evidence by assigning all studies with qualities above some relatively low threshold equal weight, in some contexts the evidence from a meta-analysis may actually provide less guidance than would a handful of the best studies on their own. Second, as we discussed at length above, experiments do not solve every problem or answer every question. Putting them at the top of an evidentiary hierarchy makes it easier to forget that they too have quality variation and can sometimes, as in the presence of substantively important general equilibrium effects, provide quite misleading policy guidance.

4.6 Cost–benefit analyses

Cost–benefit analysis exposes the full range of costs and benefits associated with a policy or program by requiring their itemisation, justification and valuation. For reasons of time and space, we do not attempt a full consideration of cost–benefit analysis here; for that we refer the reader to the vast array of journals, textbooks and conferences on the subject: see, for example, Gramlich's *Guide to Cost Benefit Analysis* (1997) or the *Journal of Benefit-Cost Analysis*. Instead, we highlight a small number of important issues often ignored in the cost–benefit analyses associated with evaluations of active labour market programs and educational interventions. Our discussion draws in part on section 10 of Heckman, LaLonde and Smith's chapter (1999).

First, we want to reiterate the importance of doing a full-blown cost–benefit analysis, especially for large programs, expensive programs and politically important programs. We do not have in mind here the sort of 'cost effectiveness' analysis that compares one program to another or one service strategy to another without a no program or no service option; in our experience these usually arise in contexts where politicians or program operators fear that the no-program or no service option will dominate the competition.

Second, we highlight the importance of considering multiple possible outcomes, including other outcomes for participating units and, as noted in the preceding section, spillovers to related units. 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 behaviours, and crime. Lechner and Wiehler (2010) find effects of German training programs on fertility. Both the original non-experimental Mathematica evaluation of the US Job Corps program, summarised by Long, Mallar, and Thornton (1981), and the more recent experimental one, summarised by Burghardt et al. (2001), stand out on this dimension, in particular for their important findings regarding the impacts of that program on participants' criminal activities. Some outcomes present real challenges to the analyst who must convert them to dollar terms, as with the primary school test scores in the Krueger cost–benefit analysis (2003) of the experimental class-size reduction in Tennessee. But, as noted in Section 3.4 in relation to general equilibrium effects, 'hard to measure' does not imply 'equals zero', despite what one might infer from reading the existing literature.

Third, a complete cost–benefit analysis should account for what economists call the deadweight costs of taxation or the marginal cost of public funds. These costs raise the social cost of one dollar of program budget to well over one dollar. They combine the direct costs of operating the tax collection system and the indirect costs imposed on society via the effects of (distortionary) taxes on behaviour. For example, income taxes lead workers to consume more leisure than they otherwise would, and so lower their utility relative to a world without income taxes. Resources spent in tax avoidance also figure into these costs. These costs will vary across countries depending on the mix of tax types (for example, income, consumption, value-added or excise) and tax rates (and perhaps local differences in behaviour conditional on these). The exact magnitude of these costs remains controversial in the scholarly literature, which suggests the wisdom of using two or three defensible values in a given cost–benefit analysis to give a sense of the sensitivity of the results to this parameter; see Dahlby's recent monograph (2008) on this topic.

Fourth, evaluations typically have available only a few years of follow-up data. For programs expected to have impacts in the medium and long term, this implies the need to project the impact estimates to time periods outside the data. In some cases, the cost–benefit performance of a program may depend critically on the persistence of impacts observed in the period covered by the available data in future periods. As such, the results of the cost–benefit analyses can be presented conditional on multiple assumptions about the persistence of any estimated program impacts.

Heckman, LaLonde and Smith (1999) provide an example of a cost–benefit analysis that does this. The assumptions about benefit persistence should build on findings on the persistence of impacts in similar programs drawn from the literature.

Fifth, 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 described by Burgess (2010).

Sixth, we both often experience a sense of wonder when we learn in response to questions about the cost of particular public programs, as we often do, that no good data exist on this score. Even some quite modestly sized businesses know their average and marginal cost structures in great detail, as they recognise the critical role these costs play in making sensible management decisions. Public managers, in contrast, often know little beyond their total agency budget and a couple of major line items, such as labour costs, and can offer only a sad face when asked about the costs associated with the marginal or average participant, let alone the costs of particular service components. Serious benefit–cost analysis requires good data on costs, data that public agencies ought, in any event, to have handy both to guide their decisionmaking and to justify their activities to the taxpaying public.

Finally, as discussed in Section 3.5, 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. Once again, a sensitivity analysis including alternative estimates of the general equilibrium effects drawn from the literature may be in order.

What can policymakers do in regard to cost–benefit analysis? To start, they can demand thorough cost–benefit analyses in those cases — new or unusual programs, expensive programs, big programs, and politically popular programs — where a cost–benefit analysis likely passes its own cost–benefit test. They can also make sure that the pieces necessary for a high-quality cost–benefit analysis get incorporated into the evaluation from the beginning, particularly in regard to the collection of data on outcomes (including longer term outcomes) and on treatment costs. They can also fund, most likely via traditional research grant programs, the work required to obtain good estimates of the marginal cost of public funds and the social discount rate.

More broadly, policymakers should face reality and accept Peter Rossi's (1987) famous 'iron law' of program evaluation, which states, 'The expected value of any net impact assessment of any large scale social program is zero.' Of course, calling it a law overstates the case to make a point, as does the quip that the US Department of Education's 'What Works Clearinghouse' should really be called the 'Nothing Works Clearinghouse.' At the same time, examples of seemingly promising treatments associated with compelling estimates of no impact litter the programmatic ground. Experimental evaluations in the United States, for example, have found no impact of the quite expensive (well over $10 000 per participant in current dollars) National Supported Work Demonstration on men (Couch 1992); no impact of youth programs under the Job Training Partnership Act (Bloom et al. 1997; Orr et al. 1996); and no impact of politically popular abstinence-only sex education curricula relative to its traditional competitors (Trenholm et al. 2007). A similar fate befell the Bush Administration's highly touted Reading First program in Abt Associates' regression-discontinuity based evaluation (Gamse et al. 2008). Policymakers should treat these results as good news, as they allow them to free up resources for promising new programs (or even to return some resources to the long-suffering taxpayer).

4.7 Alternatives to econometric program evaluation

In this section we briefly address some of the leading alternatives to serious econometric program evaluation. At the bottom we put charlatans of the sort who fill in a 'sites of oppression matrix', as described by Gregory (2000). A few steps up the ladder reside the 'guns for hire' consulting firms that cater to the crowd that knows the answer it wants in advance, as with the sorts of ex ante evaluations of professional sports facilities that rely heavily on magic multipliers; see, for example, the critique by Crompton (1995) and the papers by Noll and Zimbalist (1997).

At the top of the heap sit performance management and customer satisfaction or participant self-evaluation. The performance management 'movement' got going with Osborne and Gaebler's book *Reinventing Government* (1992). Since that time, it has exploded within the public sectors of many developed countries, including in the United States under the Clinton Administration (recall the National Performance Review) and under the Bush II administration (with its Program Assessment Rating Tool, or PART). We have no objection to many aspects of performance evaluation, such as thinking seriously about program goals, collecting good data on program inputs, outputs and outcomes or simply riding people to get them to work harder and think harder about their jobs.

The trouble comes when performance management systems confuse outcomes with impacts. For example, in many countries, something analogous to what the United States calls the 'entered employment rate' constitutes a core performance measure for active labour market programs. This rate consists of the fraction of the program's enrollees employed at some particular point (for example, 13 weeks) after leaving the program. In the terminology of the treatment effects literature, the performance measure consists of the mean of the treated outcome for the treated units. The performance measure omits any explicit counterfactual; put differently, it says nothing about what the employment rate would have been among participants had they not participated. This omission encourages, often with the help of misguided or mendacious program managers, the idea that the counterfactual equals zero, so that the outcomes summarised by such performance measures also represent impacts. Rather obviously, this interpretation encourages an overly positive view of program effectiveness. In addition, using performance measures that capture outcome levels means that high performance reflects both value-added and selection on untreated outcome levels. The literature frames this as outcome-based performance measures providing programs with an incentive to 'cream-skim' by differentially serving individuals who would have good labour market outcomes whether or not the program helps them.

In short, commonly used performance measures do not correspond to program impacts and so cannot substitute for econometric evaluation methods that do, in fact, estimate program impacts. They also provide an incentive for strategic behaviour on the part of the organisations that face them. For more on these points, as well as broader discussions of the plusses and minuses of performance management, we recommend the works of Heckman, Heinrich and Smith (2002), Barnow and Smith (2004), Radin (2006), Heinrich (2007), and Courty et al. (2010). For policymakers, the key lies in not asking performance management to do things it cannot do, like provide impact estimates.

Finally, we have what one might call participant self-evaluation. This can range from the sort of generic customer satisfaction questions used by many firms to questions that implicitly suggest some sort of counterfactual. For example, the New Chance evaluation in the United States had this question: 'Using the 0 to 10 scale, where zero is completely dissatisfied and 10 is completely satisfied, how satisfied were you overall with the New Chance program?'. The National JTPA Study had this one: 'Do you think that the training or other assistance that you got from the program helped you get a job or perform better on the job?'.

Sometimes the responses to such questions get highlighted in impact evaluations when the econometric estimates turn out badly, as if the fact that some large percentage of the customers say they liked the program makes up for a low mean impact on earnings. Recent research by Smith, Whalley and Wilcox (2010) indicates that program participants do not do a very good job of estimating impacts relative to a counterfactual, at least not with the sorts of questions presently used in evaluations. Like them, we favour additional research with alternative question designs. For the present, though, we suggest not relying on participant self-evaluations as substitutes for econometric impact estimates. We do think that participant reports have an important role to play in aspects of process evaluations, such as rating the courtesy and helpfulness of program staff.

4.8 Data quality

The quality of the underlying data plays a crucial role in determining the quality of both experimental and non-experimental evaluations. Policymakers can exert real influence here, in a very non-political, 'good government' sort of way, to improve the quality of evaluation research.

Administrative data

Administrative data increasingly form the basis of econometric program evaluations. They have a number of advantages relative to survey data. Generally, they cover long time periods, allowing the use of longitudinal methods or allowing conditioning on rich histories of outcomes and program participation in evaluations that assume selection on observed variables. They also reduce the cost of looking at longer-term impacts. Administrative data also typically, though not always, do a better job of measuring the extent and nature of treatment received than surveys; see the related discussion by Smith and Whalley (2010). Administrative data usually allow access to a whole population, which means no issues of survey non-response and larger numbers of data points for analysis than with surveys, whose higher marginal costs typically limit data collection.

At the same time, as described by Hotz and Scholz (2002), administrative data have some weaknesses for evaluation purposes — weaknesses that relatively inexpensive administrative changes can in part ameliorate. For one, administrative data often lack key covariates required in 'selection on observed variables' evaluation strategies. In some cases, the field exists in the file but remains empty for many observations and contains measurement error when filled in. Changes in software to encourage reliable data entry, along with data audits and links with other, more reliable data sets can solve these problems, and greatly improve the value of administrative data not just for impact evaluation but for process evaluation and everyday management and monitoring tasks. The main ingredient required is prioritisation of administrative data quality by policymakers, including making the effort to design institutions that allow linkages across data sets (and reasonably quick access to data) for research and evaluation purposes, while maintaining reasonable privacy protection. Policymakers can also encourage the collection of valuable data not already collected, such as the caseworker evaluations of the employability of unemployed workers routinely collected in Swiss administrative data.

Survey data

Survey data remains an important (if somewhat reduced) component of evaluation research. Surveys allow the collection of data not likely to show up in administrative data sets, including conditioning variables such as pre-program attitudes and outcome variables such as self-reported health or customer satisfaction. Survey data can also make up for problems with existing administrative data on variables such as schooling.

Policymakers can help maintain and improve the quality of survey data in three main ways. First, they can insist on response rates high enough to avoid criticisms about non-random non-response (and non-random attrition from longitudinal surveys) as well as over-reliance on statistical corrections for these problems. The US Office of Management and Budget requires response rates of 80 per cent. The leaders in the US survey industry, such as the National Opinion Research Center at Chicago and the Institute for Social Research at Michigan, routinely attain this level of response in their major research data sets. They do so despite the widespread view in the survey world that obtaining such rates has become more difficult due to lifestyle changes combined with 'survey fatigue' resulting from frequent use of surveys by commercial and advocacy groups. It just requires some money and, perhaps more importantly, the expertise. Policymakers can provide the first and hire the second.

Policymakers can also support methodological research on the potential value of new types of variables in program evaluation. This includes work on new types of conditioning variables such as the measures of risk aversion, time preference, financial knowledge, and 'ability' developed for use in the US Health and Retirement Study as well as on low cost biomarkers such as the hand traces that indicate testosterone levels based on finger length. It also includes research on how best to collect information on sensitive outcomes such as crime and sexual activity targeted by some programs. It makes sense to collect these variables first as part of large general social science data sets, then do research to determine their value, and then to add them to evaluation surveys.

4.9 Institutions

In this section we offer some ideas for relatively modest institutional changes that have the potential to improve the quality of evaluation work.

Public use data

Subject to privacy concerns, a well-documented public use data set should be one of the products ('deliverables') associated with every major evaluation. Public use (or, more accurately, researcher use) data allow independent verification of the official evaluation findings. Further, these data allow additional sensitivity analyses and the application of additional econometric methods beyond those in the official evaluation. They also encourage the production of valuable additional research, much of it of direct or indirect interest to government, at little or no cost. Academics from tenured professors to lowly graduate students will jump at the chance to work with good data when it becomes available. In addition, as any researcher doing empirical work knows, the possibility of future replication, and therefore of future public embarrassment if a mistake is found, provides a powerful motivator to thought and care.

Public use data sets exist for many major US evaluations. The Upjohn Institute for Employment Research maintains a set of evaluation datasets that it checks, documents and sells at cost to researchers. Similarly, MDRC, known for its role in many of the US welfare-to-work experiments, has a formal process to provide access to some of its evaluation datasets. The amount of knowledge about low income labour markets and about how to do econometric policy evaluation generated just from re-analysis of the data from the National Supported Work Demonstration and the National JTPA Study is simply huge, particularly when compared to the small cost of preparing the data sets for research use.

Peer review

Academia relies heavily on peer review in both the publication process and the hiring process. In our view, peer review also helps to increase the quality of program evaluations undertaken by governments. Most governments already do some of this but in many cases they could do more. We have in mind four specific avenues for increased peer review. The first consists of the use of outside experts as part of 'technical working groups' to oversee the development of an evaluation as it progresses from design, to implementation, to data collection, and finally to report writing. These should include both subject area experts and methods experts. The second consists of the presentation of evaluation results at professional meetings and conferences, ideally prior to the completion of the final report, so that comments received can affect its substance. The third consists of publication of evaluation findings in peer-reviewed academic and policy journals. Independent review by academic journals subjects technical aspects of the methods and interpretations of the official evaluation to outside scrutiny. The fourth consists of incorporating written discussant comments as part of the final evaluation report. We have in mind here what was done in Westat's evaluation of the US Employment Service and also what the *Journal of the American Statistical Association* sometimes does with important and potentially controversial articles, such as that by Heckman and Hotz (1989).

Increasing peer review improves the quality of evaluation work directly, through the comments provided by the reviewers, as well as indirectly, as the anticipation of expert scrutiny focuses and increases evaluator effort. All of these forms of additional review, particularly publication in peer-reviewed journals, have the side benefit of increasing the number of scholars, policy analysts, program managers and policymakers who learn about the methods and findings of the evaluations. This in turn should lead to increases in both the quality and quantity of related policy discussions and thereby, one imagines, to improved future policy choices. Policymakers can foster the sorts of additional review suggested here via the simple expedient of including it in the statement of work (and the evaluation budget) when commissioning evaluations from outside, or demanding it from in-house evaluators.

Encouraging interaction among academics, government and consultants

We think that more interaction among academics doing evaluation work, government evaluators and evaluation consumers, and the consultants who often produce evaluations leads to better and more useful evaluations. It also helps build a knowledgeable constituency within government for serious evaluation. This sort of interaction can take many forms, but we have personally observed the value of professional meetings such as the annual research conference operated by the Association for Public Policy and Management (APPAM, publishers of the *Journal of Policy Analysis and Management*). Consultants, as well as in-house evaluators from government agencies, get a chance to show off their work to a broader audience as well as get useful feedback. Policy oriented academics get to present their work to an audience particularly knowledgeable about the policy process and the institutions. We have also witnessed the value of having academics spend time in the government, either in roles such as the chief economist at a particular agency or in more directly research-oriented roles, wherein an economist on sabbatical might spend a year at an agency working with their data, getting writing done and interacting with the staff.

Institute of Education Sciences

Perhaps no single organisation in the United States has had a bigger effect on the quality of evaluation work in the last decade than the Institute of Education Sciences (IES). In terms of its direct effects, it has transformed the nature of federally funded evaluation of educational programs through its emphasis on funding high-quality evaluations using random assignment or regression discontinuity designs. It has brought together experts in economics, education, and other fields along with top evaluation consulting firms to conduct these evaluations. In the process it has generated valuable evidence on the effectiveness of programs such as alternative teacher certification, teacher mentoring and computer-aided mathematics instruction.

Perhaps even more important have been the indirect effects, operating through several channels. First, the IES has funded interdisciplinary training programs for education researchers at leading universities, with all of the programs having strong components in quantitative evaluation methodology and economics. Second, the IES has changed the way it runs its research grant programs to better emphasise serious quantitative research, particularly research using random assignment designs. Third, the IES has revamped the What Works Clearinghouse, a research collection, quality rating and synthesis institution modelled on the Cochrane Collaboration in medicine and the Campbell Collaboration in the social sciences, with the goal of raising the standards of empirical work in education. More details on the theory and practice behind IES can be found in the material by Rudalevige (2008) and US Institute of Education Sciences (2008). We heartily recommend it as an example both for other countries and for other policy areas.

4.10 Summary and conclusions

Evidence-based policy, and good government more generally, rest on a foundation of serious, hard-headed program evaluation. This paper has emphasised what policymakers can do to increase the quality of such program evaluation on a variety of different dimensions. The following points summarise our views and recommendations.

1. Be clear about the policy question of interest. Be sure that the econometric evaluation methods and data collection strategies adopted provide an answer to that question, even in a world where the impacts of programs vary across

persons and where both persons and program staff may make participation choices based on their informal estimates of individual impacts.

2. Use random assignment when possible. Frequent use of random assignment signals that a government is serious about evaluation and serious about basing policy on evidence. Infrequent use of random assignment sends the opposite signal. Keep in mind that randomisation can often aid in evaluation even without a no-treatment control group.
3. The success or failure of non-experimental evaluation methods depends critically on decisions about the design and implementation of the program, and on the quality of the administrative and/or survey data used in the evaluation. Thoughtful choices about program implementation and design can create useful variation in participation across time, space or persons that allows for credible evaluation. Slick econometric methods will not, other than by chance, overcome weak data or careless program design and implementation.
4. General equilibrium effects of programs matter. Analyses of these effects require different methods, in general, than analyses of the impacts of programs on their participants. Funding such analyses makes sense for large-scale programs. When a new analysis is impossible, the literature should guide an analysis of the sensitivity of the cost–benefit performance of the program to likely levels of general equilibrium effects.
5. Cost–benefit analysis represents the final step in program evaluation. Programs cost real money that taxpayers would otherwise use for their own ends. They deserve a full and complete accounting of the success or failure of the programs operated on their behalf, one that takes account the marginal cost of public funds, the possibility of general equilibrium effects, and the possibility of effects on outcomes other than those directly targeted by the program and that makes reasonable assumptions about the persistence of program impacts beyond the data.
6. Avoid the siren call of popular alternatives (such as performance management and surveys of customer satisfaction) to serious program evaluation. Both have their uses but the literature makes clear that neither provides a reliable substitute for econometric evaluations.
7. Many relatively simple and inexpensive institutional changes can have important effects on evaluation quality. These include the creation of public use data sets, greater use of outside expertise during evaluation design and execution, and publication of evaluation findings in peer-reviewed outlets as well as the creation of institutions to encourage deeper interaction between government, academics involved in evaluation research, and evaluation consultants.

References

Altonji, J., Elder, T. and Taber, C. 2005, ‘Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools’, *Journal of Political Economy,* vol. 113, no. 1, pp. 151–84.

Angelucci, M. and De Giorgi, G. 2009, ‘Indirect effects of an aid program: how do cash transfers affect ineligibles’ consumption?’, *American Economic Review,* vol. 99, no. 1, pp. 486–508.

Angrist, J. 1998, ‘Estimating the labor market impact of voluntary military service using social security data on military applicants’, *Econometrica,* vol. 66, no. 2, pp. 249–88.

—— and Evans, W. 1998, ‘Children and the parents’ labor supply: evidence from exogenous variation in family size’, *American Economic Review,* vol. 88, no. 3, pp. 450–77.

—— and Pischke, J-S. 2009, *Mostly Harmless Econometrics*, Princeton University Press, Princeton.

Banerjee, A. and Duflo, E. 2009, ‘The experimental approach to development economics’, *Annual Review of Economics,* vol. 1, pp. 151–78.

Barnow, B. 2010, Setting up social experiments: the good, the bad and the ugly, Manuscript, Johns Hopkins University, unpublished.

—— and Smith, J. 2004, ‘Performance management of U.S. job training programs: lessons from the Job Training Partnership Act’, *Public Finance and Management,* vol. 4, no. 3, pp. 247–87.

Bertrand, M., Duflo, E. and Mullainathan, S. 2004, ‘How much should we trust differences-in-differences estimates?’, *Quarterly Journal of Economics,* vol. 119, no. 1, pp. 249–75.

Bitler, M., Gelbach, J. and Hoynes, H. 2006, ‘What mean impacts miss: distributional effects of welfare reform experiments’, *American Economic Review,* vol. 96, no. 4, pp. 988–1012.

Black, D., Smith, J., Berger, M. and Noel, B. 2003, ‘Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system’, *American Economic Review,* vol. 93, no. 4, pp. 1313–27.

Blattman, C. 2008, Impact evaluation 2.0, Presentation to the Department for International Development (DFID), London, UK.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W. and Bos, J. 1997, ‘The benefits and costs of JTPA Title II-A programs: key findings from the National

Job Training Partnership Act Study', *Journal of Human Resources,* vol. 32, no. 3, pp. 549–76.

Blundell, R, Dearden, L. and Sianesi, B. 2005, 'Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey', *Journal of the Royal Statistical Society: Series A,* vol. 168, no. 3, pp. 473–512.

Bound, J., Jaeger, D. and Baker, R. 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,* vol. 90, no. 430, pp. 443–50.

Burgess, D. 2010, 'Toward a reconciliation of alternative views on the social discount rate', in Burgess, D. and Jenkins, G. (eds), *Discount Rates for the Evaluation of Public-Private Partnerships,* McGill-Queen's University Press, Montreal, pp. 131–56.

Burghardt, J., Schochet, P., McConnell, S., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J. and Jackson, R. 2001, *Does the Job Corps Work? Summary of the National Job Corps Study,* Mathematica Policy Research, Princeton, NJ.

Busso, M., DiNardo, J. and McCrary, J. 2009a, New evidence on the finite sample properties of propensity score matching and reweighting estimators, Manuscript, University of Michigan, unpublished.

——, ——, —— 2009b, Finite sample properties of semiparametric estimators of average treatment effects', Manuscript, University of Michigan, unpublished.

Caliendo, M., and Kopeinig, S. 2008, 'Some practical guidance for the implementation of propensity score matching', *Journal of Economic Surveys,* vol. 22, no. 1, pp. 31–72.

Cameron, C. and Trivedi, P. 2005, *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.

Card, D. and Krueger, A. 1994, 'Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania', *American Economic Review,* vol. 84, no. 4, pp. 772–93.

—— and Sullivan, D. 1988, 'Measuring the effect of subsidized training programs on movements in and out of employment', *Econometrica,* vol. 56, no. 3, pp. 497–530.

Cook, T. 2008, '"Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics', *Journal of Econometrics,* vol. 142, no. 2, pp. 636–54.

Couch, K. 1992, ‘New evidence on the long-term effects of employment and training programs’, *Journal of Labor Economics,* vol. 10, no. 4, pp. 380–8.

Courty, P., Heckman, J., Heinrich, C., Marschke, G. and Smith, J. 2010, *Performance Standards in a Government Bureaucracy,* W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Crompton, J. 1995, ‘Analysis of sports facilities and events: eleven sources of misapplication’, *Journal of Sports Management*, vol. 9, no. 1, pp. 14–35.

Dahlberg, M. and Forslund, A. 2005, ‘Direct displacement effects of labour market programmes’, *Scandinavian Journal of Economics,* vol. 107, no. 3, pp. 475–94.

Dahlby, B. 2008, *The Marginal Cost of Public Funds: Theory and Applications,* MIT Press, Cambridge, MA.

Davidson, C. and Woodbury, S. 1993, ‘The displacement effects of reemployment bonus programs’, *Journal of Labor Economics,* vol. 11, no. 4, pp. 575–605.

Deaton, A. 2009, ‘Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development’, NBER Working Paper no. 14690.

Dehejia, R. and Wahba, S. 1999, ‘Causal effects in nonexperimental studies: reevaluating the evaluation of training programs.’ *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1053–62.

——, —— 2002, ‘Propensity score matching methods for non-experimental causal studies’, *Review of Economics and Statistics,* vol. 84, no. 1, pp. 151–61.

De Giorgi, G. 2008, ‘Long-term effects of a mandatory multi-stage program: the New Deal for Young People in the UK’, Institute for Fiscal Studies Working Paper 05/08.

Dillon, S. 2008, ‘An initiative on reading is rated ineffective’, *New York Times*, 2 May.

Djebbari, H. and Smith, J. 2008, ‘Heterogeneous impacts in PROGRESA’, *Journal of Econometrics,* vol. 145, no. 1–2, pp. 64–80.

Dolton, P. and Smith, J. 2010, The econometric evaluation of the New Deal for Lone Parents, Manuscript, University of Michigan, unpublished.

Doolittle, F. and Traeger, L. 1990, *Implementing the National JTPA Study*, MDRC, New York.

Eckel, C.C., Johnson, C.A. and Montmarquette, C. 2005, ‘Saving decisions of the working poor: short- and long-term horizons’, in Carpenter, J., Harrison, G. and List, J. (eds), *Field Experiments in Economics: Research in Experimental Economics*, Volume 10, JAI Press, Greenwich, CT, pp. 219–60.

——. ——. —— and Rojas, C. 2007, ‘Debt aversion and the demand for loans for postsecondary education’, *Public Finance Review,* vol. 35, pp. 233–62.

Evans, W. and Kim, B. 2006, ‘Patient outcomes when hospitals experience a surge in admissions’, *Journal of Health Economics,* vol. 25, no. 2, pp. 365–88.

—— and Lein, D. 2005, ‘The benefits of prenatal care: evidence from the PAT bus strike’, *Journal of Econometrics,* vol. 125, no. 1–2, pp. 207–39.

Falk, A. and Fehr, E. 2003, ‘Why labour market experiments?’, *Labour Economics,* vol. 10, no. 4, pp. 399–406.

Frölich, M. and Lechner, M. 2010, ‘Exploiting regional treatment intensity for the evaluation of active labor market policies’, *Journal of the American Statistical Association*, forthcoming.

Gamse, B., Jacob, R.T., Horst, M., Unlu, F., Bozzi, L., Caswell, L., Rodger, C., Smith, W.C., Brigham, N. and Rosenblum, S. 2008, *Reading First Impact Study Final Report* (NCEE 2009-4038), National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, Washington, DC.

Goldberger, A. 1972a, Selection bias in evaluating treatment effects: some formal illustrations, Manuscript, University of Wisconsin, unpublished.

—— 1972b, Selection bias in evaluating treatment effects: the case of interaction, Manuscript, University of Wisconsin, unpublished.

—— 2008, ‘Selection bias in evaluating treatment effects: some formal illustrations’, in Millimet, D., Smith, J. and Vytlacil, E. (eds), *Modeling and Evaluating Treatment Effects in Economics: Advances in Econometrics,* vol. 21, pp. 1–31.

Gramlich, E. 1997, *A Guide to Benefit-Cost Analysis,* 2nd edn, Waveland Press.

Greenberg, D. and Shroder, M. 2004, *Digest of Social Experiments,* 3rd edn, Urban Institute Press, Washington, DC.

Gregory, A. 2000, ‘Problematizing participation: a critical review of approaches to participation in evaluation theory’, *Evaluation,* vol. 6, no. 2, 179–99.

Heckman, J. 1979, ‘Sample selection bias as a specification error’, *Econometrica,* vol. 47(1), pp. 153–61.

—— 1996, ‘Comment’, in Feldstein, M. and Poterba, J. (eds), *Empirical Foundations of Household Taxation*, University of Chicago Press, Chicago, pp. 32–8.

——, Heinrich, C. and Smith, J. 2002, ‘Understanding incentives in public organizations’, *Journal of Human Resources,* vol. 37, no. 4, pp. 778–811.

——, Hohmann, N., Smith, J. and Khoo, M. 2000, 'Substitution and dropout bias in social experiments: a study of an influential social experiment', *Quarterly Journal of Economics,* vol. 115, no. 2, pp. 651–94.

—— and Hotz, V.J. 1989, 'Choosing among alternative methods of evaluating the impact of social programs: the case of manpower training', *Journal of the American Statistical Association,* vol. 84, no. 408, pp. 862–74.

——, Ichimura, H., Smith, J. and Todd, P. 1998, 'Characterizing selection bias using experimental data', *Econometrica,* vol. 66, no. 5, pp. 1017–98.

—— LaLonde, R. and Smith, J. 1999, 'The economics and econometrics of active labor market programs', in Ashenfelter, O. and Card, D. (eds), *Handbook of Labor Economics,* vol. 3A, North-Holland, Amsterdam, pp. 1865–2097.

——, Lochner, L. and Taber, C. 1998, 'Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents', *Review of Economic Dynamics.* vol. 1, no. 1, pp. 1–58.

——, and Smith, J. 2000, 'The sensitivity of experimental impact estimates: evidence from the National JTPA Study', in Blanchflower, D. and Freeman, R. (eds), *Youth Employment and Joblessness in Advanced Countries*, University of Chicago Press for NBER, Chicago, pp. 331–56.

——, —— and Clements, N. 1997, 'Making the most of programme evaluations and social experiments: accounting for heterogeneity in programme impacts', *Review of Economic Studies,* vol. 64, no. 4, pp. 487–535.

——, —— and Taber, C. 1998, 'Accounting for dropouts in social experiments', *Review of Economics and Statistics,* vol. 80, no. 1, pp. 1–14.

—— Tobias, J. and Vytlacil, E. 2001, 'Four parameters of interest in the evaluation of social programs', *Southern Economic Journal,* vol. 68, no. 2, pp. 210–23.

—— Urzua, S. 2009, 'Comparing IV with structural models: what simple IV can and cannot identify', NBER Working Paper no. 14706.

Heinrich, C. 2007, 'Evidence-based policy and performance management: challenges and prospects in two parallel movements', *American Review of Public Administration,* vol. 37, no. 3, pp. 255–77.

Hirano, K., Imbens, G., Rubin, D. and Zhou, X-H. 2000, 'Assessing the effect of an influenza vaccine in an encouragement design', *Biostatistics,* vol. 1, pp. 69–88.

Hotz, V.J., Imbens, G. and Klerman, J. 2006, 'Evaluating the differential effects of alternative welfare-to-work training components: a reanalysis of the California GAIN program', *Journal of Labor Economics,* vol. 24, no. 3, pp. 521–66.

—— and Scholz, J.K. 2002, ‘Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data’ in *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press, pp. 275-315.

Ichino, A, Mealli, F. and Nannicini, T. 2008, ‘From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?’, *Journal of Applied Econometrics,* vol. 23, pp. 305–27.

Imbens, G. 2009, ‘Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009)’, NBER Working Paper no. 14896.

—— and Angrist, J. 1994, ‘Identification and estimation of local average treatment effects’, *Econometrica,* vol. 62, no. 4, pp. 467–76.

—— and Lemieux, T. 2008, ‘Regression discontinuity designs: a guide to practice’, *Journal of Econometrics,* vol. 142, no. 2, pp. 615–35.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., Ross, C., Schochet, P., and Swank, P. 2007, *National Evaluation of Early Reading First: Final Report*, US Government Printing Office, US Department of Education, Institute of Education Sciences, Washington, DC.

Kemple, J., Doolittle, F., and Wallace, J. 1993, *The National JTPA Study: Site Characteristics and Participation Patterns*. Manpower Demonstration Research Corporation, New York, NY.

Kochar, A. 1999, ‘Smoothing consumption by smoothing income: hours-of-work responses to idiosyncratic agricultural shocks in rural India’, *Review of Economics and Statistics,* vol. 81, no. 1, pp. 50–61.

Krueger, A. 2003, ‘Economic considerations and class size’, *Economic Journal,* vol. 113, no. 485, pp. F34–F63.

LaLonde, R. 1986, ‘Evaluating the econometric evaluations of training programs with experimental data’, *American Economic Review,* vol. 76, no. 4, pp. 604–20.

Lechner, M. and Smith, J. 2007, ‘What is the value added by case workers?’, *Labour Economics,* vol. 14, no. 2, pp. 135–51.

—— and Wiehler, S. 2010, ‘Kids or courses: gender differences in the effects of active labor market programs’, *Journal of Population Economics*, forthcoming.

—— and Wunsch, C. 2009, ‘Are training programs more effective when unemployment is high?’, *Journal of Labor Economics,* vol. 27, no. 4, pp. 653-92.

Lee, D. and Lemieux, T. 2009, ‘Regression discontinuity designs in economics’, NBER Working Paper no. 14723.

—— and McCrary, J. 2009, ‘The deterrence effect of prison: dynamic theory and evidence’, Manuscript, University of California, Berkeley, unpublished.

Leigh, A. 2009, What evidence should social policymakers use?, Manuscript, Australian National University, unpublished.

Lise, J., Seitz, S. and Smith, J. 2010, Equilibrium policy experiments and the evaluation of social programs, Manuscript, University of Michigan, unpublished.

Long, D., Mallar, C. and Thornton, C. 1981, ‘Evaluating the benefits and costs of the Job Corps’, *Journal of Policy Analysis and Management,* vol. 1, no. 1, pp. 55–76.

McConnell, S., Decker, P. and Perez-Johnson, I. 2006, ‘The role of counseling in voucher programs: findings from the individual training account experiment’, Manuscript, Mathematica Policy Research, unpublished.

McCrary, J. 2008, ‘Manipulation of the running variable in the regression discontinuity design: a density test’, *Journal of Econometrics,* vol. 142, no. 2, pp. 698–714.

Meyer, B. 1995, ‘Natural and quasi-experiments in economics’, *Journal of Business and Economic Statistics,* vol. 13, no. 2, pp. 151–61.

Milligan, K. and Stabile, M. 2007, ‘The integration of child tax credits and welfare: evidence from the Canadian National Child Benefit Program’, *Journal of Public Economics,* vol. 91, no. 1–2, pp. 305–26.

Moffitt, R. 1991, ‘Program evaluation with nonexperimental data’, *Evaluation Review,* vol. 15, no. 3, pp. 291–314.

Morris, P. and Michalopoulos, C. 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,* vol. 24, no. 2, pp. 20-39.

Neumark, D. and Wascher, W. 2000, ‘Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: comment’, *American Economic Review,* vol. 90, no. 5, pp. 1362–96.

Noll, R. and Zimbalist, A. 1997, *The Economic Impact of Sports Teams and Facilities*, Brookings Institution, Washington, DC.

Oreopoulos, P. 2006, ‘Estimating average and local average treatment effects of education when compulsory schooling laws really matter’, *American Economic Review,* vol. 96, no. 1, pp. 152–75.

Orr, L., Bloom, H., Bell, S., Doolittle, F., Lin, W. and Cave, G. 1996, *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study,* Urban Institute Press, Washington DC.

Osborne, D. and Gaebler, T. 1992, *Reinventing Government: How The Entrepreneurial Spirit is Transforming the Public Sector,* Perseus, Boulder, CO.

Radin, B. 2006, *Challenging the Performance Movement: Accountability, Complexity and Democratic Values*. Georgetown University Press, Washington, DC.

Raudenbusch, S. and Bryk, A. 2001, *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn, Sage, New York.

Rossi, P. 1987, The Iron Law of Evaluation and Other Metallic Rules, *Research in Social Problems and Public Policy*, no. 4, pp 3-20.

Rudalevige, A. 2008, 'Structure and science in education research', in Hess, F. (ed), *When Research Matters*, Harvard Education Press, Cambridge, MA, pp. 17–40.

Schochet, P. 2008, *Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations* (NCEE 2008-4026), National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, Washington, DC.

Smith, G. and Pell, J. 2003, 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *British Medical Journal,* vol. 327, pp. 20–7.

—— and Staghøej, J. 2010, Using statistical treatment rules for assign of participants in labor market programs, Manuscript, University of Michigan, unpublished.

—— and Todd, P. 2005, 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics,* vol. 125, no. 1–2, pp. 305–53.

—— and Whalley, A. 2010, How well do we measure public job training?, Manuscript, University of Michigan, unpublished.

——, —— and Wilcox, N. 2010, Are program participants good evaluators?, Manuscript, University of Michigan, unpublished.

Todd, P. and Wolpin, K. 2005, 'Assessing the impact of a school subsidy program in Mexico using a social experiment to validate a dynamic behavioral model of child schooling and fertility', *American Economic Review,* vol. 96, no. 5, pp. 1384–1417.

Trenholm, C., Devaney, B., Fortson, K., Quay, L., Wheeler, J and Clark, M. 2007, *Impacts of Four Title V, Section 510 Abstinence Education Programs: Final Report*, Mathematica Policy Research, Princeton, NJ.

US General Accounting Office 1996, *Job Training Partnership Act: Long-Term Earnings and Employment Outcomes* (Report HEHS-96-40), US Government Printing Office, Washington, DC.

US Institute of Education Sciences 2008, *Rigor and Relevance Redux: Director's Biennial Report to Congress* (IES 2009-6010), US Department of Education, Washington, DC.

Van der Klaauw, W. 2008, 'Regression-discontinuity analysis: a survey of recent developments in economics', *Labour: Review of Labour Economics and Industrial Relations,* vol. 22, no. 2, pp. 219–45.

Wooldridge, J. 2002, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.