Journal of Economic Literature 48 (June 2010): 399–423
http:www.aeaweb.org/articles.php?doi=10.1257/jel.48.2.399

Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)

Guido W. Imbens*

Two recent papers, Deaton (2009) and Heckman and Urzua (2009), argue against what they see as an excessive and inappropriate use of experimental and quasi-experimental methods in empirical work in economics in the last decade. They specifically question the increased use of instrumental variables and natural experiments in labor economics and of randomized experiments in development economics. In these comments, I will make the case that this move toward shoring up the internal validity of estimates, and toward clarifying the description of the population these estimates are relevant for, has been important and beneficial in increasing the credibility of empirical work in economics. I also address some other concerns raised by the Deaton and Heckman–Urzua papers. (JEL C21, C31*)*

1. *Introduction*

Two recent papers, Angus S. Deaton (2009) (Deaton from hereon) and James J. Heckman and Sergio Urzua (2009) (HU from hereon), argue against what they see as an excessive and inappropriate use of experimental and quasi-experimental methods in empirical work in economics in the last decade.[1] Deaton and HU reserve much of their scorn for the local average treatment effect (LATE) introduced in the econometric literature by Guido W. Imbens and Joshua D. Angrist (1994) (IA from hereon). HU write: "Problems of identification and interpretation are swept under the rug and replaced by 'an effect' identified by IV that is often very difficult to interpret as an answer to an interesting economic question" (HU, p. 20). Deaton writes: "This goes beyond the old story of looking [for] an object where the light is strong enough to see; rather, we have control over the light, but choose to let it fall where

* Imbens: Harvard University. I have benefitted from discussions with Joshua Angrist, Susan Athey, Abhijit Banerjee, David Card, Gary Chamberlain, Esther Duflo, Kei Hirano, Geert Ridder, Chuck Manski, Sendhil Mullainathan, and Jeffrey Wooldridge, although they bear no responsibility for any of the views expressed here. Financial support for this research was generously provided through NSF grants 0631252 and 0820361.

[1] The papers make similar arguments, perhaps not surprisingly given Deaton's acknowledgement that "much of what I have to say is a recapitulation of his [Heckman's] arguments" (Deaton, p. 4).

it may, and then proclaim that whatever it illuminates is what we were looking for all along" (Deaton, p. 10) and "The LATE may, or may not, be a parameter of interest . . . and in general, there is no reason to suppose that it will be . . . I find it hard to make any sense of the LATE" (Deaton, p. 10).[2] Deaton also rails against the perceived laziness of these researchers by raising the "futility of trying to avoid thinking about how and why things work," (Deaton, p. 14).[3] HU wonder whether these researchers are of the opinion: "that disguising identification problems by a statistical procedure is preferable to an honest discussion of the limits of the data?" (HU, p. 19).

The fact that two such distinguished economists so forcefully[4] question trends in current practice may suggest to those not familiar with this literature that it is going seriously awry. In these comments, I will argue that, to the contrary, empirical work is much more credible as a result of the natural experiments revolution started by David Card, Angrist, Alan B. Krueger, and others in the late 1980s. Starting in the late 1980s, their work, and more recently that by development economists such as Abhijit V. Banerjee, Esther Duflo, and Michael Kremer arguing in favor of randomized experiments, has had a profound influence on empirical work. By emphasizing internal validity and study design, this literature has shown the importance of looking for clear and exogenous sources of variation in potential causes. In contrast to what Deaton and HU suggest, this issue of data quality and study design is distinct from the choice between more or less structural or theory driven models and estimation methods. In fact, recognizing this distinction, there has been much interesting work exploring the benefits of randomization for identification, estimation, and assessment of structural models. For an early example, see Jerry A. Hausman and David A. Wise (1979), who estimate a model for attrition with data from a randomized income maintenance experiment, and, for recent examples, see, among others, Card and Dean R. Hyslop (2005), who estimate a structural model of welfare participation using experimental data from Canada; Petra E. Todd and Kenneth Wolpin (2003), who analyze data from Mexico's Progresa program; Imbens, Donald B. Rubin and Bruce I. Sacerdote (2001), who estimate dynamic labor supply models exploiting random variation in unearned earnings using data from lottery winners; Duflo, Rema Hanna, and Stephen Ryan (2007), who look at the effect of randomized monitoring and financial incentives on teacher's absences; Susan Athey, Jonathan Levin, and Enrique Seira (forthcoming), who use randomized assignment of auction formats to estimate structural models of bidding behavior; Raj Chetty, Adam Looney, and Kory Kroft (2009), who look at the effect of tax saliency using experimental evidence from supermarket pricing policies; and Chetty and Emmanuel Saez (2009), who exploit random variation in information about the tax code. There is much room for such work where experimental variation is used to improve the identification and credibility of the structural models. It would put at risk the progress made in improving the credibility of empirical work in economics if this message got lost in discussions about the relative merits of structural work versus work less directly connected to economic theory or in minor squabbles about second-order technicalities such as

[2] Somewhat incongruously, Deaton views Heckman's local instrumental variables methods much more positively as the "appropriate" (Deaton, p. 14) response to dealing with heterogeneity, although the marginal treatment effect that is the cornerstone of this approach is nothing more than the limit version of the LATE in the presence of continuous instruments, e.g., Heckman, Urzua, and Edward Vytlacil (2006), see also Angrist, Kathryn Graddy, and Imbens 2000).

[3] Curiously, Deaton exempts the leaders of this movement from these charges, by declaring them "too talented to be bound by their own methodological prescriptions" (Deaton, p. 4).

[4] Deaton dismisses arguments of those he disagrees with as merely "rhetoric" no less than six times in his paper.

adjustments for heteroskedasticity in the calculation of standard errors and the Behrens–Fisher problem (e.g., Deaton, p. 33).[5]

In my view, it is helpful to separate the discussion regarding the merits of the recent literature on experiments and natural experiments into two parts. The first part concerns the questions of interest and the second the choice of methods conditional on the question. In my opinion, the main concern with the current trend toward credible causal inference in general, and toward randomized experiments in particular, is that it may lead researchers to avoid questions where randomization is difficult, or even conceptually impossible, and natural experiments are not available. There are many such questions and many of them are of great importance. Questions concerning the causal effects of macroeconomic policies can rarely be settled by randomized experiments.[6] The effect of mergers and acquisitions cannot be studied using experiments. Similarly, questions involving general equilibrium effects cannot be answered by simple experiments. In other examples, randomized experiments raise ethical concerns and are ultimately not feasible. These are not new concerns and I am sympathetic with the comments in this regard made by, for example, Dani Rodrik (2008). There is clearly much room for nonexperimental work and history abounds with examples where causality has ultimately found general acceptance without any experimental evidence. The most famous example is perhaps the correlation between smoking and lung cancer. The interpretation of this correlation as evidence of a causal effect of smoking on lung cancer is now generally accepted, without any direct experimental evidence to support it. It would be unfortunate if the current interest in credible causal inference, by insisting on sometimes unattainable standards of internal validity, leads researchers to avoid such questions. At the same time, the long road toward general acceptance of the causal interpretation of the smoking and lung cancer correlation (including Fisher's longtime skepticism about the causal interpretation of this correlation) shows the difficulties in gaining acceptance for causal claims without randomization.

However, the importance of questions for which randomization is difficult or infeasible should not take away from the fact that, for answering the questions they are designed for, randomized experiments, and other (what Card calls) design-based strategies, have many advantages. Specifically, conditional on the question of interest being one for which randomized experiments are feasible, randomized experiments are superior to all other designs in terms of credibility. Deaton's view that "experiments have no special ability to produce more credible knowledge than other methods" (Deaton, abstract) runs counter to the opinions of many researchers who have considered these issues previously. David A. Freedman, hailed by Deaton himself as "one of its [the world's] greatest statisticians"[7] (Deaton, title page, acknowledgement) is unambiguous in his opening sentence, "Experiments offer more reliable evidence on causation than observational studies" (Freedman 2006, abstract). Edward E. Leamer (1983), in his influential criticism of the state of empirical work in the 1970s, writes, "There is therefore a sharp difference between inference from randomized experiments and inference from natural

[5] Moreover, there is nothing in these issues that makes observational studies less vulnerable to them.

[6] Although, for an interesting macroeconomic study in the spirit of the modern causal literature, see Christina D. Romer and David H. Romer (2004), who study the effects of monetary policy decisions on the macroeconomy, exploiting variation in Federal Reserve policies at times when markets viewed the Federal Reserve decisions as unpredictable and, thus, essentially as random.

[7] I certainly have no disagreement with this qualification: see my endorsement on the back cover of Freedman (2010).

experiments" (Leamer, p. 33).[8] That is not to say that one may not choose to do an observational study for other reasons, e.g., financial costs, or ethical considerations, even in settings where randomized experiments are feasible. However, no other design will have the credibility that a randomized experiment would have. Suppose we are interested in a question that can be addressed by randomized experiments, for example, whether a job training program has an effect on labor market outcomes or whether class size affects educational outcomes. In such settings, the evidence from a randomized experiment is unambiguously superior to that from observational studies. As a result, randomized experiments have often been very influential in shaping policy debates, e.g., the 1965 Perry Preschool Project on early childhood interventions (see Constance Holden 1990 and Charles F. Manski 1997 for some recent discussions), the National Supported Work Demonstration experiments on labor market programs (e.g., Robert J. LaLonde 1986), or Project STAR on class size reductions (e.g., Krueger 1999). More generally, and this is really the key point, in a situation where one has control over the assignment mechanism, there is little to gain, and much to lose, by giving up this control through allowing individuals to choose their own treatment regime. Randomization ensures exogeneity of key variables, where, in a corresponding observational study, one would have to worry about their potential endogeneity.

In these comments, I will make five points from the perspective of an econometrician who is interested in, and has been involved in, the methodological aspects of this literature. First, I will give a different characterization of goals and focus of the literature Deaton and HU take issue with. For its emphasis on obtaining credible causal estimates and for developing a clear understanding of the nature of the variation that gives these estimates credibility, I will refer to this as the causal literature. Second, I will discuss briefly the origins of this causal literature, which takes its motivation partially from the failure of specific structural models, such as the Heckman selection model (e.g., Heckman 1978), to satisfactorily address endogeneity issues in the context of estimation of causal effects of labor market programs. This was famously documented by LaLonde (1986) (see also Thomas Fraker and Rebecca Maynard 1987). Third, I will elaborate on the point that, in cases where the focus is establishing the existence of causal effects and where experiments are feasible, experiments are unambiguously the preferred approach: since Ronald A. Fisher (1925) it has formally been established that randomization gives such designs a credibility unmatched by any other research design.

Fourth, I will make the case that a key contribution of the recent theoretical literature on causality has been to clarify the merits, as well as the limitations, of instrumental variables, local average treatment effects, and regression discontinuity designs in settings with heterogenous causal effects. Far from "disguising identification problems by a statistical procedure" (HU, p. 19), it was shown by IA that, in instrumental variables settings with heterogenous effects, instrumental variables methods do identify the average treatment effect for a well defined subpopulation (the compliers in the terminology from Angrist, Imbens and Rubin 1996), indexed by the instrument.[9] Although, in many cases these, what are now known as local average

[8] By natural experiments Leamer here refers to studies without formal randomization, that is, observational studies (personal communication).

[9] Although Deaton credits Heckman (1997) with establishing that in the presence of heterogenous effects the probability limit of instrumental variables estimators depends on the instrument, this was previously shown in Imbens and Angrist (1994).

treatment effects or LATEs, and similarly the estimands in regression discontinuity designs, are *not* the average effects that researchers set out to estimate, the internal validity of those estimands is often much higher than that of other estimands. I will also take issue with the Deaton and HU view that somehow instrumental variables methods are atheoretical. The exclusion and monotonicity restrictions that underlie such methods are motivated by subject matter, that is economic, rather than statistical, knowledge. Moreover, the focus on instrumental variables estimands, rather than on reduced form correlations between outcomes and exogenous variables (including instruments), is motivated by the belief that the former are more likely to be structural than the latter.[10]

In the fifth point, I discuss issues related to external validity, that is, the ability of the estimands to generalize to other populations and settings. The causal literature has emphasized internal validity over external validity, with the view that a credible estimate of the average effect for a subpopulation is preferred to an estimate of the average for the target population with little credibility. This is consistent with the biomedical literature. Although the primacy of internal validity over external validity has been criticized often in that literature, there is little support for moving toward a system where studies with low internal validity receive much weight in policy decisions. External validity is generally a more substantial problem in economics than in biomedical settings, with considerable variation in both preferences and constraints between individuals, as well as variation over time. Understanding heterogeneity in treatment effects is therefore of great importance in these settings and it has received considerable attention in the theoretical evaluation literature (see Imbens and Jeffrey M. Wooldridge 2009 for a survey) and in the experimental literature (e.g., Banerjee and Duflo 2009).

2 *Causal Models and Design-Based Approaches*

The literature that does not conform to the Deaton and HU standards of structural work is variously referred to, in a somewhat pejorative manner, as atheoretical or statistical (as opposed to economic).[11] These are not terms commonly used in this literature itself. They are also at odds with their historical use.[12] The almost complete lack of instrumental variables methods in the statistical literature makes that label an unusual one for the literature that Deaton and HU focus on in their criticism. What is shared by this literature is not so much a lack of theoretical or economic motivation but rather an explicit emphasis on credibly estimating causal effects, a recognition of the heterogeneity in these effects, clarity in the identifying assumptions, and a concern about endogeneity of choices and the role study design plays. I will therefore refer to this interchangeably as the causal or design-based literature. Early influential examples include the Card (1990) study of the impact of immigration using the Mariel Boatlift, Angrist's (1990) study of the effect of veteran status on earnings using the Vietnam era draft lottery as an instrument, and the Angrist and Krueger (1991) study

[10] "Structural" is used here in the Arthur S. Goldberger (1991) sense of invariant across populations.

[11] This literature is also sometimes referred to as "reduced form," again a misnomer. In the classical, Cowles Commission, simultaneous equations setting, the term reduced form is used to refer to the regression of the endogenous variables on the full set of exogenous variables (which is typically estimated by ordinary least squares). Equations estimated by instrumental variables methods are, in this terminology, referred to as structural equations.

[12] In an even more remarkable attempt to shape the debate by changing terminology, Deaton proposes to redefine the term "exogeneity" in such a way that "Even random numbers—the ultimate external variables—may be endogenous" (Deaton, p. 13).

of the effect of education on earnings using variation in educational achievement related to compulsory schooling laws. More recently, this has led to many studies using regression discontinuity designs (see David S. Lee and Thomas Lemieux 2010 for a review). The recent work in development economics has taken the emphasis on internal validity even further, stressing formal randomization as a systematic and robust approach to obtaining credible causal effects (see Duflo, Rachel Glennerster, and Kremer 2008 for an overview of this literature). This has led to a spectacular increase in experimental evaluations in development economics (see, for example, the many experiments run by researchers associated with the Poverty Action Lab at MIT), and in many other areas in economics, e.g., Marianne Bertrand and Sendhil Mullainathan (2004), Duflo and Saez (2003), Chetty, Looney, and Kroft (2009), and many others.

Often the focus in this literature is on causal effects of binary interventions or treatments. See Imbens and Wooldridge (2009) for a recent review of the methodological part of this literature. For example, one may focus on the effect of universal exposure to the treatment, that is, the average treatment effect, or on the effect of exposure for those currently exposed, the average effect on the treated. Even if these interventions do not directly correspond to plausible future policies, they are often useful summary statistics for such policies and, therefore, viewed as quantities of interest. A major concern in this literature is that simple comparisons between economic agents in the various regimes are often not credible as estimates of the average effects of interest because of the potential selection bias that may result from the assignment to a particular regime being partly the result of choices by optimizing agents. As a consequence, great care is applied to the problem of finding credible sources of exogenous variation in the receipt of the intervention of interest, often in combination with the innovative collection of original data sources, to remove any selection bias.

To focus the discussion, let me introduce a specific example. Suppose a state, say California, is considering reducing class size in first through fourth grade by 10 percent. Entering in the California policymakers' decision is the comparison of the cost of such a class size reduction with its benefits. Suppose that the policymakers have accurate information regarding the cost of the program but are unsure about the benefits. Ultimately the hope may be that such a reduction would improve labor market prospects of the students, but let us suppose that the state views the program as worthwhile if it improves some measure of skills, say measured as a combination of test scores, by some amount. What is the relevance for this decision of the various estimates available in the literature? Let us consider some of the studies of the effect of class size on educational outcomes. There is a wide range of such studies but let me focus on a few. First, there is experimental evidence from the Tennessee STAR experiments starting in 1985 (e.g., Krueger 1999). Second, there are estimates based on regression discontinuity designs using Israeli data (Angrist and Victor Lavy 1999). Third, there are estimates exploiting natural variation in class size arising from natural variation in cohort size using data from Connecticut reported in Caroline M. Hoxby (2000). None of these estimates directly answers the question facing the decisionmakers in California. So, are any of these three studies useful for informing our California policymaker? In my view, all three are. In all three cases, finding positive effects of class size reductions on test scores would move my prior beliefs on the effect in California toward bigger effects. Exactly how much each of the three studies would change my

prior beliefs would depend on the external and internal validity of the three studies. Specifically, the external validity of each study would depend on (i) its timing relative to the target program, with older studies receiving less weight, (ii) differences between the study population and the California target population, including the targeted grade levels in each study, and (iii) differences between the study outcomes and the goals of the California programs. In terms of these external validity criteria, the Hoxby study with Connecticut data would probably do best. In terms of internal validity, that is, of the estimate having a credible causal interpretation, the Krueger study using experimental Tennessee data would definitely, and, next, the Angrist–Lavy study with Israeli data might, do better. The main point, though, is that all three studies are in my view useful. None of the three answers directly the question of interest but the combination is considerably better than any single one. We could clearly do better if we designed a study especially to study the California question. Ideally we would run an experiment in California itself, which, five years later, might give us a much more reliable answer but it would not help the policymakers at this moment very much. If we did an observational study in California, however, I would still put some weight on the Connecticut, Tennessee, and Israeli studies. One may go further in formalizing the decision process in this case and I will do so in section 6.

Reiterating the main point, having a variety of estimates, with a range of populations and a range of identification strategies, can be useful to policymakers even if none of the individual studies directly answers the policy question of interest. It is of course unrealistic to expect that the California policymakers would be able to pick a single study from the literature in order to get an answer to a question that had not actually been posed yet when these studies were conducted. This is, again, not a new point. The proponents of randomization in the new development economics have argued persuasively in favor of doing multiple experiments (Duflo 2004; Banerjee 2007; Banerjee and Duflo 2009). It is obvious that, as Deaton comments, simply repeating the same experiment would not be very informative. However, conducting experiments on a variety of settings, including different populations and different economic circumstances, would be. As Deaton suggests, informing these settings by economic theory, much as the original negative income tax experiments were, would clearly improve our understanding of the processes as well as our ability to inform public policy.

The focus of the causal literature has been on shoring up the internal validity of the estimates and on clarifying the nature of the population these estimates are relevant for. This is where instrumental variables, local average treatment effects, and regression discontinuity methods come in. These often do not answer exactly the question of interest but provide estimates of causal effects for well-defined subpopulations under weaker assumptions than those required for identification of the effects of primary interest. As a result, a single estimate is unlikely to provide a definitive and comprehensive basis for informing policy. Rather, the combination of several such studies, based on different populations and in different settings, can give guidance on the nature of interventions that work.

Let me mention one more example. Deaton cites a study by Banerjee et al. (2007) who find differences in average effects between randomized evaluations of the same program in two locations. Banerjee et al. surmise that these differences are related to differential initial reading abilities. Deaton dismisses this conclusion as not justified by the randomization because such a question was not part of

the original protocol and would therefore be subject to data mining issues. This is formally correct, but it is precisely the attempt to understand differences in the results of past experiments that leads to further research and motivates subsequent experiments, thus building a better understanding of the heterogeneity in the effects that can assist in informing policy. See Card, Jochen Kluve, and Andrea Weber (2009) for another example of such a meta analysis and section 6 for additional discussion.

3. *LaLonde (1986): The Failure of Nonexperimental Methods to Replicate Experimental Evaluations of Labor Market Programs*

Surprisingly, neither Deaton nor HU discuss in much detail the origins of the resurgence of interest in randomized and natural experiments, and the concern with the internal validity of some of the structural modeling. HU vaguely reference the "practical difficulty in identifying, and precisely estimating, the full array of structural parameters" (HU, p. 2), but mention only, without a formal reference, a paper by Hausman (presumably Hausman 1981) as one of the papers that according to HU "fueled the flight of many empirical economists from structural models" (HU, p. 2, footnote 6). I think the origins behind this flight are not quite as obscure or haphazard as may appear from reading Deaton and HU. Neither of them mentions the role played by LaLonde's landmark 1986 paper, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data."[13] In his 1986 paper, widely cited and still widely taught in labor and econometrics courses in economics PhD programs, LaLonde studies the ability of a number of econometric methods, including Heckman's selection models, to replicate the results from an experimental evaluation of a labor market program on the basis of nonexperimental data. He concluded that they could not do so systematically. LaLonde's evidence, and subsequent studies with similar conclusions, e.g., Fraker and Maynard (1987), had a profound impact in the economics literature and even played a role in influencing Congress to mandate experimental evaluations for many federally funded programs.

It would appear to be uncontroversial that the focus in LaLonde's study, the average effect of the Nationally Supported Work (NSW) program, meets Deaton's criterion of being "useful for policy or understanding," (Deaton, abstract). The most direct evidence that it meets this criterion is the willingness of policymakers to provide substantial funds for credible evaluations of similar labor market and educational programs. Nevertheless, the question remains whether evaluation methods other than those considered by LaLonde would have led to better results. There is some evidence that matching methods would have done better. See the influential paper by Rajeev H. Dehejia and Sadek Wahba (1999), although this is still disputed, see, e.g., Jeffrey A. Smith and Todd (2005) and Dehejia (2005). See Imbens and Wooldridge (2009) for a recent review of such methods. Matching methods, however, hardly meet Deaton's criteria for "analysis of models derived from economic theory" (Deaton, p. 2). Until there are more successful attempts to replicate experimental results, it would therefore seem inescapable that there is a substantial role to be played by experimental evaluations in this literature if we want data analyses to meet Leamer's standard of being taken seriously by other researchers.

[13] At some level, LaLonde's paper makes the same point as Leamer did in his 1983 paper in which he criticized the state of empirical work based on observational studies. See the symposium in the Spring 2010 *Journal of Economic Perspectives* for a recent discussion on the effects of Leamer's criticism on subsequent empirical work.

4. *The Benefits of Randomized Experiments*

One of the most curious discussions in Deaton concerns the merits of randomized experiments. He writes: "I argue that evidence from randomized controlled trials has no special priority . . . Randomized controlled trials cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence" (Deaton, p. 4) and "Actual randomization faces similar problems as quasi-randomization, notwithstanding rhetoric to the contrary" (Deaton, abstract). These are remarkable statements. If true, in the unqualified way Deaton states them, it would throw serious doubt on the Food and Drug Administration's (FDA) insistence on randomized evaluations of new drugs and treatments. But of course Deaton's statements are wrong. Deaton is both formally wrong and wrong in spirit. Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top. Again, this is not merely my view: see the earlier quotes from Freedman and Leamer for similar sentiments.

For support for his position that "Randomization is not a gold standard" (Deaton, p. 4), Deaton quotes Nancy Cartwright (2007) as claiming that "there is no gold standard" (Cartwright 2007, quoted in Deaton, p. 4). It is useful to give a slightly longer quote from Cartwright (2007) to put her claim in perspective: "The claims of randomized controlled trials (RCTs) to be the gold standard rest on the fact that the ideal RCT is a deductive method: if the assumptions of the test are met, a positive result implies the appropriate causal conclusion. This is a feature that RCT's share with a variety of other methods, which thus have equal claim to being a gold standard" (Cartwright 2007, abstract). I agree with Cartwright that many methods have the feature that if their assumptions are met, the causal conclusions follow. However, I strongly disagree with her claim that this is what gives randomized experiments their credibility. It is not the *assumption* of randomization but the actual *act* of randomization that allows for precise quantifications of uncertainty, and this is what gives randomization a unique status among study designs. Constance Reid (1982, p. 45) quotes Jerzey Neyman concerning the importance of the actual randomization, and the attribution of this insight to Fisher: ". . . the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fishers's achievements." It is interesting to see that Cartwright does not mention Fisher's or Neyman's views on these issues in the discussion of her claims.

Far from being merely rhetoric as Deaton claims, the physical act of randomization is what allows the researcher to precisely quantify the uncertainty associated with the evidence for an effect of a treatment, as shown originally by Fisher (1925). Specifically, it allows for the calculation of exact p-values of sharp null hypotheses. These p-values are free of assumptions on distributions of outcomes, free from assumptions on the sampling process, and even free of assumptions on interactions between units and of assumptions on compliance behavior, solely relying on randomization and a sharp null hypothesis. No other design allows for this. Now this is strictly speaking a very narrow result, with the extensions to more interesting questions somewhat subtle. We can establish the uncertainty regarding the existence of a causal effect through the calculation of p-values but we cannot establish properties of estimators for, say, the average effect without additional assumptions and approximations. Unless we rule out interactions between individuals, the average effect of the treatment depends on assignments to other individuals and, thus, needs to be defined carefully. In

the absence of interactions, we can estimate the average effect without bias but the validity of confidence intervals still relies on large sample approximations (e.g., Neyman 1923; Freedman 2008). Nevertheless, even if experiments rely on some assumptions or large sample approximations for inference on average treatment effects, they do so to a lesser extent than observational studies by not requiring assumptions on the assignment mechanism.

Deaton himself hedges his remarkable claims by adding that "actual experiments are frequently subject to practical problems that undermine any claims to statistical or epistemic superiority" (Deaton, abstract), a somewhat confusing statement given that according to the earlier quotes in his view there is no initial superiority to undermine. It is true that violations of assignment protocols, missing data, and other practical problems can create complications in the analyses of data from randomized experiments. There is no evidence, however, that giving up control of the assignment mechanism and conducting an observational study improves these matters. Moreover, the suggestion that any complication, such as a violation of the assignment protocol, leads to analyses that lose all credibility accorded to randomized experiments is wrong. Again, it is both formally wrong and wrong in substance. That this suggestion is formally wrong is easiest illustrated in an example.

Consider a randomized experiment with N individuals, M randomly assigned to the treatment group and the remaining $N - M$ assigned to the control group. In the absence of complications such as noncompliance, interactions between units, and missing data, we can calculate the p-value associated with the null hypothesis of no effect of the treatment and we can also estimate the average effect of the treatment without bias, both based on the randomization distribution. Now suppose that there is noncompliance. Some individuals assigned to the treatment were not exposed to the treatment, and some individuals assigned to the control group were in fact exposed to the treatment. We can still assess the null hypothesis of no effect of the treatment using the same analysis as before as long as we take care to use the randomization distribution of the assignment to treatment rather than the receipt of treatment. There is no complication in this analysis arising from the noncompliance. The noncompliance does, however, compromise our ability to find an estimator that is unbiased for the average treatment effect. However, if, for example, the outcome is binary, we can still derive, in the spirit of the work by Manski (1990, 1995, 2003), a range of values consistent with the average treatment effect that is valid without assumptions on the compliance behavior. These bounds may be informative, depending on the data, and, in particular if the rate of noncompliance is low, will lead to a narrow range. In the presence of missing data, both the derivation of p-values and estimators will now lead to ranges of values without additional assumptions. An important role is played here by Manski's insight that identification is not a matter of all or nothing. Thus, some of the benefits of randomization formally remain even in the presence of practical complications such as noncompliance and missing data.

In his paper, Deaton also questions what we learn from experiments: "One immediate consequence of this derivation is a fact that is often quoted by critics of RCTs but is often ignored by practitioners, at least in economics: RCTs are informative about the *mean* of the treatment effects, $Y_{i1} - Y_{i0}$, but do not identify other features of the distribution. For example, the median of the difference is not the difference in medians, so an RCT is not, by itself, informative about the median treatment effect, something that could be of as much interest to policymakers as the mean treatment effect" (Deaton, p. 26). He further stresses this point by writing "Put differently, the trial might reveal an average positive

effect although nearly all of the population is hurt with a few receiving very large benefits, a situation that cannot be revealed by the RCT, although it might be disastrous if implemented" (Deaton, p. 27). These statements are formally wrong in the claims about the information content of randomized experiments and misleading in their suggestion about what is of interest to policymakers.

Regarding the first claim, here is a simple counterexample, similar to one discussed in the Heckman and Smith (1995) paper cited by Deaton. Suppose we have a randomized experiment with binary outcomes. Assume that among the controls and treated the outcome distributions are binomial with mean p_0 and p_1 respectively. If the difference $p_1 - p_0$ exceeds $1/2$, one can infer that the median effect of the treatment is one. In general, however, it is correct that the evidence from randomized experiments regarding the joint distribution of the pair $(Y(0), Y(1))$ is limited. Nevertheless, there is more information regarding quantiles than Deaton suggests. In the presence of covariates, experiments are directly informative concerning the two conditional distributions $f(Y(0)|X)$ and $f(Y(1)|X)$, and together these may lead to more informative bounds on, say, quantiles of the distribution of the difference $Y(1) - Y(0)$ than simply the two marginal distributions $f(Y(0))$ and $f(Y(1))$.

The more important issue is the second claim in the Deaton quote, that the median could be of as much interest to policymakers as the mean treatment effect or, more generally, that it is the joint distribution we are interested in, beyond the two marginal distributions. In many cases, average effects of (functions of) outcomes are indeed what is of interest to policymakers, *not* quantiles of differences in potential outcomes. The key insight is an economic one—a social planner, maximizing a welfare function that depends on the distribution of outcomes in each state of the world, would only care about the two marginal distributions, not about the distribution of the difference. Suppose that the planner's choice is between two programs. In that case, the social planner would look at the welfare given the marginal distribution of outcomes induced by the first program and compare that to the welfare given the marginal outcome distribution induced by the second program, and *not* at the joint distribution of outcomes. My argument against Deaton's claim that policymakers could be as much interested in the median treatment effect as in the mean treatment effect is not novel. As Manski (1996) writes, "Thus, a planner maximizing a conventional social welfare function wants to learn $P[Y(1)]$ and $P[Y(0)]$, not $P[Y(1) - Y(0)]$" (Manski 1996, p. 714). (Here $P[Y(w)]$ denotes the distribution of $Y(w)$.) The implication is that the planner's decision may depend on the median of the marginal distributions of $Y_i(0)$ and $Y_i(1)$ but would, in general, not depend on the median of the treatment effect $Y_i(1) - Y_i(0)$. To make this specific, let us return to Deaton's example of a program with few reaping large benefits and many suffering small losses. Manski's social planner would compare the distribution given the treatment, $P[Y(1)]$, with the distribution in the absence of the treatment, $P[Y(0)]$. The comparison would not necessarily be based simply on means but might take into account measures of dispersion, and so avoid the potential disasters Deaton is concerned about.

Deaton also raises issues concerning the manner in which data from randomized experiments are analyzed in practice. Consider a carefully designed randomized experiment with covariates present that were not taken into account in the randomization.[14] Deaton raises three issues. The

[14] In fact, one should take these into account in the design because one would always, even in small samples, be at least as well off by stratifying on these covariates as by ignoring them, e.g., Imbens et al. (2009), but that is a different matter.

first concerns inference or, more specifically, estimation of standard errors. The second is concerned with finite sample biases. The third issue deals with specification search and the exploration of multiple hypotheses. I will address each in turn. Before doing so, let me make two general comments. First, in my view, the three issues Deaton raises are decidedly second order ones. That is, second order relative to the first order issues of selection and endogeneity in observational evaluation studies that have long been highlighted in the econometric literature, prominently in work by Heckman (e.g., Heckman 1978; Heckman and Richard Robb 1985). Second, Deaton appears to be of the view that the only way experiments should be analyzed is based on randomization inference.[15] Randomization inference is still relatively rare in economics (the few examples include Bertrand, Duflo, and Mullainathan 2002 and Alberto Abadie, Alexis Diamond, and Jens Hainmueller forthcoming) and, although personally I am strongly in favor of its use (see the discussion in Imbens and Wooldridge 2009), it is not the only mode of inference even if one has experimental data. If one uses model-based inference, including regression methods, there are still well established benefits from randomized assignment of the treatment even if there are no longer exact finite sample results (e.g., Rubin 1978, 1990).[16] As I wrote in the introduction to this paper, when one is estimating a structural model it is still helpful to have experimental data. Although regression estimators are generally not unbiased under the randomization distribution, regression estimators are made more robust and credible by randomization because at least some of the assumptions underlying regression analyses are now satisfied by design.

Now let me turn to the first issue raised by Deaton, concerning the standard errors. This is an issue even in large samples. If the average effect is estimated as the difference in means by treatment status, the appropriate variance, validated by the randomization, is the robust one, allowing for heteroskedasticity (e.g., Friedhelm Eicker 1967; Peter J. Huber 1967; Halbert White 1980). Using the standard ordinary least squares variance based on homoskedasticity leads to confidence intervals that are not necessarily justified even in large samples. This point is correct and, in practice, it is certainly recommended to use the robust variance here, at least in sufficiently large samples.[17] Moreover, the standard error issue that is often the most serious concern in practice, clustering, is nowadays routinely taken into account. See Duflo, Glennerster, and Kremer (2008) for more discussion.

The second issue concerns finite sample issues. Researchers often analyze randomized experiments using regression methods, including as regressors both the treatment indicator and covariates not affected by the

[15] Under randomization inference, properties such as bias and variance are calculated over the distribution induced by random assignment for the fixed population, keeping potential outcomes with and without treatment and covariates fixed, and reassigning only the treatment indicator. This contrasts with the model-based repeated sampling perspective often used in econometrics where the covariates and the treatment indicator are fixed in repeated samples and the unobserved component in the regression function is redrawn from its distribution. See Paul R. Rosenbaum (1995) for a general discussion.

[16] Freedman (2006) argues for randomization inference whenever a randomized experiment is conducted: "Experiments should be analyzed as experiments, not as observational studies" (Freedman, abstract). I have some sympathy for that view, although that does not take away from the fact that, if one wants to estimate a structural model, one would still benefit from having experimental data.

[17] There are further complications in small samples. The most commonly used version of robust standard errors performs worse than homoskedastic standard errors in small samples. There are improvements available in the literature, especially for the simple case where we compare two sample averages, which probably deserve more attention. See, among others, Henry Scheffe (1970).

treatment. If only the treatment indicator is included in the specification of the regression function, the least squares estimator is identical to the difference in average outcomes by treatment status. As shown originally by Neyman, this estimator is unbiased, in the finite sample, over the distribution induced by randomizing the treatment assignment. As Freedman (2008) points out, if one includes additional covariates in the specification of the regression function, the least squares estimator is no longer exactly unbiased, where again the distribution is that induced by the randomization.[18] On the other hand, including covariates can substantially improve the precision if these covariates are good predictors of the outcomes with or without the treatment. In finite samples, there is therefore a tradeoff between some finite sample bias, and large sample precision gains. In practice including some covariates that are a priori believed to be substantially correlated with the outcomes, is likely to improve the expected squared error. An additional point is that if the regression model is saturated, e.g., with a binary covariate including both the covariate and the interaction of the covariate and the treatment indicator, there is no bias, even in finite samples.[19]

The third issue Deaton raises concerns the exploration of multiple specifications, for example through the estimation of average effects for various subgroups. This is formally correct, and I would certainly encourage researchers to follow more closely the protocols established by the FDA, which, for example, insists on listing the analyses to be conducted prior to the collection of the data. Again there is of course nothing specific to randomized experiments in this arguments: any time a researcher uses pretesting or estimates multiple versions of a statistical model there should be concern that the final confidence intervals no longer have the nominal coverage rate. See Leamer (1978) for a general discussion of these issues. However, I think that this is again a second order issue in the context of the comparison between randomized experiments and observational studies. In randomized experiments, one typically finds, as in LaLonde (1986), that the results from a range of estimators and specifications are robust. Had Deaton added a real example of a case where results based on experiments were sensitive to these issues, his argument would have been more convincing.

Ultimately, and this is really the key point of this section, it seems difficult to argue that, in a setting where it is possible to carry out a randomized experiment, one would ever benefit from giving up control over the assignment mechanism by allowing individuals to choose

[18] This result may come as a surprise to some researchers, so let me make it explicit in a simple example. Suppose there are three units, with covariate values $X_1 = 0$, $X_2 = 1$, and $X_3 = 2$. If assigned to the treatment, the outcomes for the three units are $Y_1(1)$, $Y_2(1)$, and $Y_3(1)$ and, if assigned to the control treatment, the outcomes are $Y_1(0)$, $Y_2(0)$, and $Y_3(0)$. The average treatment effect is $\tau = (Y_1(1) + Y_2(1) + Y_3(1))/3 - (Y_1(0) + Y_2(0) + Y_3(0))/3$. Suppose the experiment assigns one of the three units to the treatment and the other two units to the control group. Thus, there are three possible values for the assignment vector, $\mathbf{W} \in \{(1,0,0),(0,1,0),(0,0,1)\}$. Under each value of the assignment, we can calculate the value of the two estimators. The first estimator is equal to the difference in the average outcomes for the treated units and the average outcomes for the control units. The second estimator is based on the least squares regression of the observed outcome on a constant, the treatment indicator W_i, and the covariate X_i. For example, if $\mathbf{W} = (0,0,1)$, the first estimator is equal to $\hat{\tau}_{dif} = Y_3(1) - Y_2(0)/2 - Y_1(0)/2$ and the second estimator is equal to $\hat{\tau}_{ols} = Y_3(1) - 2Y_2(0) + Y_1(0)$. It is simple to calculate the expectation of these two estimators over the randomization distribution. For the difference estimator, the expectation is equal to the average treatment effect τ but, for the least squares estimator, the expectation is equal to $(Y_1(1) + Y_2(1) + Y_3(1))/3 - (-Y_1(0)/2 + 4Y_2(0) - Y_3(0)/2)/3$, which in general differs from τ. This bias disappears at rate $1/N$ as the sample size increases, as shown in Freedman (2006).

[19] A separate issue is that it is difficult to see how finite sample concerns could be used as an argument against actually doing experiments. There are few observational settings for which we have exact finite sample results.

their own treatment status.[20] In other words, conditional on the question, the methodological case for randomized experiments is unassailable and widely recognized as such, and none of the arguments advanced by Deaton and HU weaken that. I do not want to say that, in practice, randomized experiments are generally perfect or that their implementation cannot be improved, but I do want to make the claim that giving up control over the assignment process is unlikely to improve matters. It is telling that neither Deaton nor HU give a specific example where an observational study did improve, or would have improved, on a randomized experiment, conditional on the question lending itself to a randomized experiment.

5. *Instrumental Variables, Local Average Treatment Effects, and Regression Discontinuity Designs*

In some settings, a randomized experiment would have been feasible, or at least conceivable, but was not actually conducted. This may have been the result of ethical considerations, or because there was no particularly compelling reason to conduct an experiment, or simply practical reasons. In some of those cases, credible evaluations can be based on instrumental variables or regression discontinuity strategies. As a rule, such evaluations are second best to randomized experiments for two reasons. First, they rely on additional assumptions and, second, they have less external validity. Often, however, such evaluations are all we have. The theoretical econometrics literature in the last two decades has clarified what we can learn, and under what conditions, about the intervention in those settings.[21] In doing so, this literature has made many connections to the statistics and psychology literature on observational studies. Rather than leading to "unnecessary rhetorical barriers between disciplines working on the same problems" (Deaton, p. 2), this has been a remarkably effective two-way exchange, leading to substantial convergence in the statistics and econometrics literatures, both in terms of terminology and in the exchange of ideas. On the one hand, economists have now generally adopted Rubin's potential outcome framework (Rubin 1974, 1990; Rosenbaum and Rubin 1983), labeled the Rubin Causal Model by Paul W. Holland (1986), which formulates causal questions as comparisons of unit-level potential outcomes.[22] Although this potential outcome framework is a substantial departure from the Cowles Commission general set up of simultaneous equations models, it is closely related to the interpretation of structural equations in, for example, Trygve Haavelmo (1943).[23] On the other hand, statisticians gained an appreciation for, and understanding of, instrumental variables methods. See, for example, what is probably the first use of instrumental

[20] Of course it is possible that the question of interest itself involves the choice of treatment status. For example, if we are interested in a job training program that would be implemented as a voluntary program, the experimental design should involve randomization of the option to enroll in the program, and not randomization of enrollment itself.

[21] For a recent review of this literature, see Imbens and Wooldridge (2009).

[22] Compare, for example, the set up in Heckman and Robb (1985), which predates the potential outcome set up, with that in Heckman (1990), which adopts that framework.

[23] Jan Tinbergen distinguishes between "any imaginable price" π that enters into the demand and supply functions and the "actual price" p that is determined by the market clearing in his notation. Subsequently this distinction between potential outcomes and realized outcomes has become blurred. With the work of the Cowles Foundation, the standard textbook notation directly relates a matrix of observed endogenous outcomes ($\mathbf{Y}$) to a matrix of exogenous variables ($\mathbf{X}$) and a matrix of unobserved "residuals" ($\mathbf{U}$), linked by a set of unknown parameters (B and Γ):

$$\mathbf{Y}\Gamma + \mathbf{X}B = \mathbf{U},$$

representing causal, structural relationships. This notation has obscured many of the fundamental issues and continues to be an impediment to communication with other disciplines.

variables published in the mainstream medical literature, although still written by economists, Mark McClellan, Barbara J. McNeil, and Joseph P. Newhouse (1994). Special cases of these methods had been used previously in the biostatistics literature, in particular in settings of randomized experiments with one-sided compliance (e.g., M. Zelen 1979), but no links to the econometrics literature had been made. Furthermore, economists have significantly generalized applicability and understanding of regression discontinuity designs (Jinyong Hahn, Todd, and Wilbert van der Klaauw 2001; Justin McCrary 2008; Lee 2008; Imbens and Karthik Kalyanaraman 2008) and now in turn influence the psychology literature where these designs originated. See William R. Shadish, Thomas D. Cook, and Donald T. Campbell (2000) and Cook (2008) for a historical perspective. Within economics, however, the results in IA and Hahn, Todd, and van der Klaauw (2001) are unusual ("the opposite of standard statistical practice" Deaton, p. 9). As a consequence, these papers have generated a substantial degree of controversy as echoed in the quotes from Deaton and HU. Let me offer some comments on this.

The standard approach in econometrics is to state precisely what the object of interest is at the outset of an analysis. Let me use Angrist's (1990) famous draft lottery study as an example. In that case, one may be interested in the average causal effect of serving in the military on earnings. Now suppose one is concerned that simple comparisons between veterans and nonveterans are not credible as estimates of average causal effects because of selection biases arising from unobserved differences between veterans and nonveterans. Let us consider the arguments advanced by Angrist in support of using the draft lottery number as an instrument. The first key assumption is that draft eligibility is exogenous. Since it was actually randomly assigned, this is true by design in this case.

The second is that there is no direct effect of the instrument, the lottery number, on the outcome. This is what Angrist, Imbens, and Rubin (1996) call the exclusion restriction.[24] This is a substantive assumption that may well be violated. See Angrist (1990) and Angrist, Imbens, and Rubin (1996) for discussions of potential violations.[25]

The third assumption is what IA call monotonicity, which requires that any man who would serve if not draft eligible, would also serve if draft eligible.[26] In this setting, monotonicity, or as it is sometimes called "no-defiers," seems a very reasonable assumption. Although Deaton quotes Freedman as wondering "just why are there no defiers" (Freedman 2006, quoted in Deaton, p. 37) and Heckman and Urzua (2010) write about "arbitrary conditions like 'monotonicity' that are central to LATE" (Heckman and Urzua 2010, p. 8), the monotonicity assumption is often well motivated from the perspective of optimizing agents. Increasing the value of the instrument, in the draft lottery example corresponding to giving the person a lottery number that implies the person will be more likely to be subject to the draft, raises the cost of staying out of the military. It would seem reasonable to assume that the response to this increase in costs, for each optimizing individual, is an increased likelihood of serving in the military. This interpretation of changes in the value of the instrument corresponding to increases in the net benefits of receiving

[24] Deaton actually calls this second assumption "exogeneity" in an unnecessary and confusing change from conventional terminology that leads him to argue that even random numbers can fail to be exogenous.

[25] For example, the extension of formal schooling to avoid the draft could lead to violations of the exclusion restriction.

[26] In another unnecessary attempt to change established terminology, HU argue that this should be called "uniformity."

the treatment does not hold in all cases and, when IA introduced the assumption, they discuss settings where it need not be plausible. Nevertheless, it is far from an arbitrary assumption and often plausible in settings with optimizing agents. In addition, Angrist, Imbens, and Rubin discuss the implications of violations of this assumption.

These three assumptions are not sufficient to identify the average effect of serving in the military for the full population. However, as shown by IA, these assumptions are sufficient to identify the average effect on the subpopulation of what Angrist, Imbens, and Rubin (1996) call compliers, the local average treatment effect or LATE. Compliers in this context are individuals who were induced by the draft lottery to serve in the military, as opposed to never-takers who would not serve irrespective of their lottery number, and always-takers, who would volunteer irrespective of their lottery number. But, Deaton might protest, this is not what we said we were interested in! That may be correct, depending on what is the policy question. One could imagine that the policy interest is in compensating those who were involuntarily taxed by the draft, in which case the compliers are exactly the population of interest. If, on the other hand, the question concerns future drafts that may be more universal than the Vietnam era one, the overall population may be closer to the population of interest. In that case, there are two alternatives that do focus on the average effect for the full population. Let me briefly discuss both in order to motivate the case for reporting the local average treatment effect. See also Manski (1996) for a discussion of these issues.

One principled approach is Manski's (1990, 1996, 2003) bounds, or partial identification, approach. Manski might argue that one should maintain the focus on the overall average effect and derive the bounds on this estimand given the assumptions one is willing to make. Manski's is a coherent perspective and a useful one. While I have no disagreement with the case for reporting the bounds on the overall average treatment effect, there is, in my view, a strong case for also reporting estimates for the subpopulation for which one can identify the average effect of interest, that is the local average treatment effect. The motivation for this is that there may be cases with wide bounds on the population average effect, some of which are, and some of which are not, informative about the presence of any effects. Consider an example of a randomized evaluation of a drug on survival, with one-sided noncompliance and with the randomized assignment as an instrument for receipt of treatment. Suppose the bounds for the average effect of the treatment are equal to $[-3/16, 5/16]$. This can be consistent with a substantial negative average effect for compliers, lowering survival rates by 1/4, or with a substantial positive average effect for compliers, raising survival rates by 1/4.[27] One would think that, in the first case, a decisionmaker would be considerably less likely to implement universal adoption of the treatment than in the second, and so reporting only the bounds might leave out relevant information.

A second alternative approach to the focus on the local average treatment effect is to complement the three assumptions that allowed for identification of the average effect for compliers, with additional assumptions that allow one to infer the overall average effect, at least in large samples. The concern is that the assumptions that allow

[27] To be specific, let the probability of complier and never-takers be equal to 1/2. With the endogenous regressor (receipt of treatment) denoted by X_i and the instrument (assignment of treatment) denoted by Z_i, let $p_{zx} = \text{pr}(Y = 1 | X = x, Z = z)$. In the first example, $p_{00} = 1/4$, $p_{10} = 1/8$, and $p_{11} = 1/8$. In the second example, $\tilde{p}_{00} = 1/2$, $\tilde{p}_{10} = 5/8$, and $\tilde{p}_{11} = 5/8$. In both cases the sharp bounds on the average treatment effect are $[-3/16, 5/16]$, in the first example $\tau_{late} = -1/4$, and in the second example $\tilde{\tau}_{late} = 1/4$.

one to carry out this extrapolation are of a very different nature from, and may be less credible than, those that identify the local average treatment effect. For that reason, I would prefer to keep those assumptions separate and report both the local average treatment effect, with its high degree of internal but possibly limited external validity, and possibly add a set of estimates for the overall average effect with the corresponding additional assumptions, with lower internal, but higher external, validity. Let us be more specific in the context of the Angrist study. One might write down a model for the outcome (e.g., earnings) denoted by Y_i, depending on veteran status V_i:

$$Y_i = \alpha + \beta \cdot V_i + \varepsilon_i.$$

In addition, one might write down a Heckman-style latent index model (Heckman 1978, 1990) for the decision to serve in the military as a function of the instrument Z_i (draft eligibility):

$$V_i^* = \pi_0 + \pi_1 \cdot Z_i + \eta_i.$$

The latent index V_i^* represents the difference in utility from serving, versus not serving, in the military with the observed veteran status V_i equal to

$$V_i = \begin{cases} 1 & \text{if } V_i^* > 0, \\ 0 & \text{if } V_i^* \leq 0. \end{cases}$$

The inclusion of the instrument Z_i in the utility function can be thought of as reflecting the cost a low lottery number imposes on the action of not serving in the military. Suppose that the only way to stay out of the military if drafted is through medical exemptions. In that case, it may well be plausible that the instrument is valid. Health status is captured by the unobserved component η_i: individuals in poor health $\eta_i < -\pi_0 - \pi_1$ (never-takers in the AIR terminology) would not serve even if drafted, individuals with $-\pi_0 - \pi \leq \eta_i < -\pi_0$ (compliers) would serve if drafted but not as volunteers, and individuals with $-\pi_0 \leq \eta_i$ (always-takers) would always serve. Note that this model implies the monotonicity or no-defiers condition, although, unlike in the IA set up, the assumption is implicit, rather than explicit.

Although not widely used anymore, this type of model was very popular in the 1980s, as one of the first generation of models that explicitly took into account selection bias (Heckman 1978, 1990). Note that this model embodies all the substantive assumptions underlying the local average treatment effect. Thus, the instrumental variables estimator can be justified by reference to this, admittedly simple, structural model.

Although originally this type of model was often used with a distributional assumption (typically joint normality of (η_i, ε_i)), this is not essential in this version of the model. Without any distributional assumptions, only assuming independence of ε_i and Z_i is sufficient for identifying the average effect of military service, β. More important is the assumption of a constant effect of veteran status. Such an assumption is rarely implied by theory and is often implausible on substantive grounds (e.g., with binary outcomes). Suppose we relax the model and explicitly allow for heterogenous effects:

$$Y_i = \alpha + (\beta + \nu_i) \cdot V_i + \varepsilon_i,$$

where ν_i captures the heterogeneity in the effect of veteran status for individual i. If we maintain joint normality (now of the triple $(\varepsilon_i, \eta_i, \eta_i)$), we can still identify the parameters of the model, including β, that is, the average effect of veteran status. See, for example, Anders Björklund and Robert Moffitt (1987). Unlike in the constant effect model, however, in this case the normality assumption is not innocuous. As Heckman (1990) shows, a nonparametric version of

this model is not identified unless the probability of veteran status, as a function of the instrument Z_i, is arbitrarily close to zero and one for some choices of the instrument. As this is implied by the range of the instrument being unbounded, this is often referred to as "identification at infinity" (Gary Chamberlain 1986; HU). In the case with a binary instrument, this assumption is easy to verify. In the Angrist study, the probability of serving in the military for the draft eligible and noneligible is far from zero and one, and so nonparametric identification arguments based on identification-at-infinity fail. The key contribution of IA was the insight that, although one could not identify the average effect for the overall population, one could still identify the average effect for compliers, or the LATE.[28] In the structural model above, compliers are the individuals with $\pi_0 - \pi_1 \leq \eta_i < \pi_0$. Think again of the case where the never-takers with $\eta_i < -\pi_0 - \pi_1$ correspond to individuals in poor health. These individuals cannot be induced to serve in the military through the draft. It seems intuitively clear that we cannot identify the average effect of military service for this group from such data because we never see them serving in the military. So, the problem in this case is not so much that researchers are "trying to avoid thinking about how and why things work" (Deaton, p. 14) but that there is little basis for credible extrapolation from the local average treatment effect to the overall average effect.

Reporting the local average treatment effect solely or in combination with bounds or point estimates for the overall average based on additional assumptions is, thus, emphatically *not* motivated by a claim that the local average treatment effect is the sole or even primary effect of interest. Rather, it is motivated by a sober assessment that estimates for other subpopulations do not have the same internal validity and by an attempt to clarify what can be learned from the data in the absence of identification of the population average effect. It is based on a realization that, because of heterogeneity in responses, instrumental variables estimates are a distinct second best to randomized experiments. Let me end this discussion with a final comment on the substantive importance of what we learn in such settings. Although we do not learn what the average effect is of veteran status, we can, in sufficiently large samples, learn for a particular, well-defined subpopulation, what the effect is. We may then wish to extrapolate to other subpopulations, even if only qualitatively, but given that the nature of those extrapolations is often substantially less credible than the inferences for the particular subpopulation, it may be useful to keep these extrapolations separate from the identification of the effect for compliers.

These arguments are even more relevant for the regression discontinuity case. In the sharp regression discontinuity case, we learn about the average effect of a treatment at a fixed value of the covariate. Let us consider Jordan D. Matsudaira's (2008) example of the effect of summer school attendance on subsequent school performance. Matsudaira uses comparisons of students just above and just below the threshold on the test score that leads to mandatory summer school attendance. Students close to this margin are likely to be different from those far away from the margin. At the same time, there is no reason to think that only students at the margin are of interest: the effect of summer school on students with test scores far below the margin is likely to be of considerable

[28] Althought fifteen years after its introduction Deaton still finds it hard to make sense of the LATE, Heckman, in at least some of his work, appears to value this contribution, writing "It is a great virtue of the LATE parameter that it makes the investigator stick to the data at hand, and separate out the aspects of an estimation that require out of sample extrapolation or theorizing from aspects of an estimation that are based on observable data" (Heckman 1999, p. 832).

interest as well but, in the absence of credible models for extrapolation, there may be no credible estimates for that group.

Fuzzy regression discontinuity designs rank even lower in terms of external validity. As pointed out by Hahn, Todd, and van der Klaauw (2001) in arguably the most important contribution of economists to the regression discontinuity design literature, fuzzy regression discontinuity designs combine the limitations of sharp regression discontinuity designs, in that they only refer to units with a particular value of the covariates, with those of instrumental variables estimates, in that they only reflect on compliers. However, for this subpopulation, these designs often have great internal validity. Many convincing examples have now been published. See the survey paper by Lee and Lemieux (2009) and the special issue of the journal of econometrics (Imbens and Lemieux 2008). Again, researchers do not necessarily set out to estimate the average for these particular subpopulations but, in the face of the lack of internal validity of estimates for other subpopulations, they justifiably choose to report estimates for them.

6. *Internal versus External Validity*

Much of the debate between structural and causal approaches ultimately centers on the weight researchers put on external validity versus internal validity of a study. To be precise, by a study I have in mind a combination of a population, a causal effect of interest, and an estimator. By internal validity I mean the credibility of the estimator as an estimator of the causal effect of interest, and by external validity I mean the generalizability of the causal effect to other populations.[29] The concern is typically that randomized experiments may do well in terms of internal validity but poorly in terms of external validity, relative to structural models.[30] There is no disagreement that both internal and external validity are important. See Banerjee and Duflo (2009) for a recent discussion in the context of experimental evaluations in development economics. Returning to the class size example from section 2, Angrist and Lavy (1999), Hoxby (2000), and Krueger (1999) do not study the effect of class size as a historical phenomenon: they want to inform the policy debate on class size. Similarly, Card (1990) is presumably not interested in solely in the effect of the Mariel Boatlift, rather he is interested in informing the debate on the effects of immigration of low-skilled workers. In order to be useful in informing policy, a study needs to have internal validity (have a credible causal interpretation for the population it refers to) as well as external validity (be relevant for the populations the treatment may be extended to). In many disciplines, the weights placed on different studies are heavily loaded in favor of internal validity. The FDA insists on elaborate protocols to ensure the internal validity of estimates, with much less emphasis on their external validity. This has led, at times, to the approval of drugs with a subsequent reversal of that decision after the drug was found

[29] This is in line, with, for example, Shadish, Cook, and Campbell (2002), who define internal validity as "the validity of inferences about whether observed covariation . . . reflects a causal relationship," and external validity as "the validity of inferences about whether the cause–effect relationship holds over variation in persons, settings, treatment variables, and measurement variables." It also agrees with Rosenbaum (2010) who writes "A randomized experiment is said to have a high level of 'internal validity' in the sense that the randomization provides a strong or 'reasoned' basis for inference about the effects of the treatment . . . on the . . . individuals in the experiment," and "'External' validity refers to the effects of the treatment on people not included in the experiment."

[30] Although Cartwright (2007) surprisingly has the opposite view: "Despite the claims of RCTs [randomized clinical trials] to be the gold standard, economic models have all the advantages when it comes to internal validity" and "But it seems that RCTs have the advantage over economic models with respect to external validity" (Cartwright 2007, p. 19).

to have adverse effects on populations that were underrepresented in the original study populations. Part of this is unavoidable. First, legally, randomized experiments can only be conducted with informed consent by participants and there is no systematic method for ensuring that the population of those who consent is representative of the population of interest. Second, after a successful randomized experiment, the target population may well change. If a treatment is in a randomized trial demonstrated to be beneficial for moderately sick patients, physicians may well be tempted to use it for sicker patients that were not part of the original study. Doing a second experiment on a population of sicker patients would not always be an option and would not be ethical if the first trial on the population of moderately sick individuals showed a substantial beneficial effect of the treatment. Third, other things may change between the experiment and the subsequent adoption that affects the efficacy of the treatment. Again, this is unavoidable in practice.

In economic applications, the issue of external validity is considerably more severe. In many biomedical treatments the effects are through relatively stable biological mechanisms that often generalize readily to other populations. A vaccine for a particular strain of HIV that prevents infection in the United States has a high likelihood of working for the same strain in Africa as well. In contrast, an educational reform that is found to raise test scores in England is unlikely to be directly applicable to the United States given the differences in educational institutions and practices.

It may be helpful to put some more structure on this problem.[31] Suppose we have a number of units. To be specific, I will refer to them as states. We are interested in the effect of an intervention, e.g., putting a price cap into place at p_1 versus at p_0, on demand for a particular commodity in a particular state, say California. For ease of exposition, let us assume that $p_1 - p_0 = 1$. Let the expected difference in demand, at the two potential values for the price cap, be denoted by θ_s, indexed by state s. States may differ in the expected effect θ_s because they differ in terms of institutions or because they differ in terms of population composition. Let us denote the relevant characteristics of the states by X_s and, for purposes of this discussion, let us assume we observe X_s.

Now suppose we have a structural economic model for the household level demand function:

$$D_i = \beta_0 + \beta_1 \cdot p + \beta_2 \cdot I_i \cdot p + \varepsilon_i,$$

where D_i is household level demand, I_i is household income, and ε_i are unobserved differences between households. The parameters β are structural parameters, common to all states (structural in the Goldberger 1991 sense of being invariant to changes in the population). Given this model, the difference in expected demand in state s if the price is fixed at p_1 versus p_0 is

$$\begin{aligned}\theta_s &= E[D_i \mid S_i = s, P_i = p_1] \\ &\quad - E[D_i \mid S_i = s, Pi = p_0] \\ &= \beta_1 + \beta_2 \cdot E[I \mid S = s].\end{aligned}$$

Let $X_s = E[I \mid S = s]$ be average income in state s, so that we can write

$$\theta_s = g(X_s, \beta) = \beta_1 + \beta_2 \cdot X_s.$$

We are interested solely in the difference in average outcomes in California,

$$\theta_{ca} = g(X_{ca}, \beta) = \beta_0 + \beta_1 \cdot X_{ca}.$$

[31] This discussion is partly based on conversations with Banerjee and Mullainathan.

Now suppose we have data from an experiment in Tennessee, where randomly selected individuals were faced with a price of p_1, and others with a price of p_0. Thus, with a sufficiently large sample, we would learn from the Tennessee experiment the value of $\theta_{tn} = g(X_{tn}, \beta)$.

Suppose we also have data from an observational study from Connecticut. In this state, we have a random sample of demand, income, and prices, (D_i, I_i, P_i), for $i = 1, \ldots N$. We may be concerned that in this state prices are endogenous, and so let us assume that we also observe an instrument for price, Z_i. If the instrument is valid, and conditional on income it is both correlated with prices and uncorrelated with ε_i, this will allow us to estimate the structural parameters β using two-stage least squares. Let us allow for the possibility that the instrument is not valid, or more generally for misspecification in the structural model. In that case, $\hat{\beta}_{ct}$, the estimator for β based on Connecticut data, need not be consistent for β. Let us denote the probability limit of the estimator by β_{ct}—we index this probability limit by the state to capture the possibility that if the same structural model was estimated in a different state, the bias might well be different.

The first question now is how we would choose between two estimates of the intervention in California: the experimental one from Tennessee,

$$\hat{\theta}^{exp}_{ca} = \theta_{tn},$$

versus the structural one, based on parameter estimates from Connecticut, combined with the characteristics from California,

$$\hat{\theta}^{struct}_{ca} = g(X_{ca}, \beta_{ct}).$$

In principle, the choice between the two estimators would depend on the variation in effect θ_s and in the variation in the pseudo-structural parameter β_s. In the absence of additional information, one may need to rely on prior beliefs. If one believes there is little variation in θ_s, one might prefer $\hat{\theta}^{exp}_{ca}$. If one believed the structural model was close to correctly specified, one would prefer $\hat{\theta}^{struct}_{ca}$. Note the benefits in this case of experimental data: if the structural model had actually been estimated on experimental data, there would be no bias and β_{ct} would be equal to β and, thus, $g(X_{ca}, \beta_{ct})$ would be equal to θ_{ca}. That is not always the case. If the structural model was richer, a simple experiment with randomly assigned prices would not necessarily pin down all structural parameters. However, in general, it will help pin down some combination of the structural parameters by forcing the model to fit the experimental evidence. This is closely related to the sufficient statistics approach in Chetty (2009).

The answer to the first question may also differ if the experiment in Tennessee focused on a question that differed from that in California. If the experiment in Tennessee involved randomly assigning prices of p_2 and p_3, rather than the price levels that enter into the California question, p_0 and p_1, it may be difficult to estimate θ_{ca} from the Tennessee results. This would not pose any conceptual problems from the structural model perspective.

A second question is what one would do if one had both the experimental evidence from Tennessee and the observational data from Connecticut. In that case, one could, in the spirit of the LaLonde (1986) evaluation of econometric evaluation methods, compare the experimental estimate for Tennessee, θ_{tn}, with the structural one based on Connecticut estimates, $\hat{\theta}^{struct}_{tn} = g(X_{tn}, \beta_{ct})$. The comparison of θ_{tn} and $\hat{\theta}^{struct}_{tn}$ reflects on the adequacy of the structural model. If the structural model passes the test, there is a stronger case for using the structural model to predict the effect of the intervention in California. If the prediction fails, however, the conclusion

is that the structural model is not adequate and, thus, invalidates $\hat{\theta}_{ca}^{struct}$. This test does not reflect in any way on the experimental estimate $\hat{\theta}_{ca}^{exp}$.

A third question concerns the information content of additional experiments. With two or more experiments we would be able to update our beliefs on the amount of variation in θ_s. It obviously would not help much if we did the second experiment in a state very similar to Tennessee but, if we did the second experiment in a state very different from Tennessee and ideally more similar to California, we would likely learn much about the amount of variation in θ_s. If we have detailed information on X_s, having a substantial number of experiments may enable us to approximate the function $g(x;\beta)$ without directly estimating β, simply fitting a flexible functional form to $E[\theta_s|X_s] = g(X_s;\gamma)$. If we can approximate this function accurately, we would be able to predict the effect of the intervention in California. In this case, one could also incorporate different experiments, e.g., those involving other price caps. If there is any choice, one should do the experiments in a wide range of settings, that is, in the current example, in states with different X_s. The analyses by Card, Kluve, and Weber (2009), V. Joseph Hotz, Imbens, and Julie H. Mortimer (2005), Kremer and Alaka Holla (2008), and Raghabendra Chattopadhyay and Duflo (2004) fit into this framework.

The fourth question concerns the benefits of multiple observational studies. This is not quite so clear. In many cases, one would expect that repeated observational studies in different locations would have similar biases generated through similar selection mechanisms. Finding that multiple observational studies lead to the same results is therefore not necessarily informative. To get a handle on the bias, the difference $\beta_s - \beta$, we would need observational study from states that do not have the same biases as the first state, Connecticut. Identifying such states may be more difficult than finding a state with potentially different effects θ_s: it may well be that the biases in observational studies would be similar in all states, arising from the same selection mechanisms. Rosenbaum (1987) discusses similar issues arising in the presence of multiple control groups in observational studies.

7. *Conclusion*

Deaton offers a critical appraisal of the methodologies currently in fashion in development economics. He argues that randomized experiments have no special role in the hierarchy of evidence and, as do Heckman and Urzua, argues somewhat presumptuously that instrumental variables methods do not answer interesting questions. He suggests moving toward more theory-based studies and away from randomized and natural experiments. In these comments, I take issue with some of these positions and caution against his recommendations. The causal or design-based literature, going back to the work in labor economics by Angrist, Card, Krueger, and others, and the current experimental literature in development economics, including work by Duflo, Banerjee, and Kremer, has greatly improved the standards of empirical work by emphasizing internal validity and clarifying the nature of identifying assumptions. Although it would be regrettable if this trend led researchers to avoid questions that cannot be answered through randomized or natural experiments, it is important not to lose track of the great strides made by this literature toward improving the credibility of empirical work.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. Forthcoming. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*.

Angrist, Joshua D. 1990. "Lifetime Earnings and the

Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–36.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies*, 67(3): 499–527.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434): 444–55.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106(4): 979–1014.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics*, 114(2): 533–75.

Athey, Susan, Jonathan Levin, and Enrique Seira. Forthcoming. "Comparing Open and Sealed Bid Auctions: Theory and Evidence from Timber Auctions." *Quarterly Journal of Economics*.

Banerjee, Abhijit V. 2007. *Making Aid Work*. Cambridge and London: MIT Press.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, 122(3): 1235–64.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics*, 1: 151–78.

Banerjee, Abhijit V., and Ruimin He. 2008. "Making Aid Work." In *Reinventing Foreign Aid*, ed. William R. Easterly, 47–92. Cambridge and London: MIT Press.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2002. "How Much Should We Trust Differences-in-Differences Estimates?" National Bureau of Economic Research Working Paper 8841.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991–1013.

Björklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *Review of Economics and Statistics*, 69(1): 42–49.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43(2): 245–57.

Card, David, and Dean R. Hyslop. 2005. "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers." *Econometrica*, 73(6): 1723–70.

Card, David, Jochen Kluve, and Andrea Weber. 2009. "Active Labor Market Policy Evaluations: A Meta-analysis." Institute for the Study of Labor Discussion Paper 4002.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 84(4): 772–93.

Cartwright, Nancy. 2007. "Are RCTs the Gold Standard?" *BioSocieties*, 2(1): 11–20.

Chamberlain, Gary. 1986. "Asymptotic Efficiency in Semi-parametric Models with Censoring." *Journal of Econometrics*, 32(2): 189–218.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica*, 72(5): 1409–43.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics*, 1: 451–88.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review*, 99(4): 1145–77.

Chetty, Raj, and Emmanuel Saez. 2009. "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients." National Bureau of Economic Research Working Paper 14836.

Cook, Thomas D. 2008. "'Waiting for Life to Arrive': A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics." *Journal of Econometrics*, 142(2): 636–54.

Deaton, Angus S. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." National Bureau of Economic Research Working Paper 14690.

Dehejia, Rajeev H. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics*, 125(1–2): 355–64.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, 94(448): 1053–62.

Duflo, Esther. 2004. "Scaling Up and Evaluation." In *Annual World Bank Conference on Development Economics, 2004: Accelerating Development*, ed. François Bourguignon and Boris Pleskovic, 341–69. Washington, D.C.: World Bank; Oxford and New York: Oxford University Press.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics, Volume 4*, ed. T. Paul Schultz and John Strauss, 3895–3962. Amsterdam and San Diego: Elsevier, North-Holland.

Duflo, Esther, Rema Hanna, and Stephen Ryan. 2007. "Monitoring Works: Getting Teachers to Come to School." Unpublished.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics*, 118(3): 815–42.

Eicker, Friedhelm. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." In *Proceedings of the Berkeley Symposium on Mathematical*

Statistics and Probability, Volume 1, 59–82. Berkeley: University of California Press.
Fisher, Ronald A. 1925. *The Design of Experiments*, First edition. London: Oliver and Boyd.
Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, 22(2): 194–227.
Freedman, David A. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review*, 30(6): 691–713.
Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics*, 40(2): 180–93.
Freedman, David A. 2010. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, ed. David Collier, Jasjeet Sekhon, and Philip B. Stark. Cambridge and New York: Cambridge University Press.
Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, Mass. and London: Harvard University Press.
Haavelmo, Trygve. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica*, 11(1): 1–12.
Hahn, Jinyong, Petra E. Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201–09.
Hausman, Jerry A. 1981. "Labor Supply." In *How Taxes Affect Economic Behavior*, ed. Henry J. Aaron and Joseph A. Pechman, 27–72. Washington, D.C.: Brookings Institution Press.
Hausman, Jerry A., and David A. Wise. 1979. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica*, 47(2): 455–73.
Heckman, James J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica*, 46(4): 931–59.
Heckman, James J. 1990. "Varieties of Selection Bias." *American Economic Review*, 80(2): 313–18.
Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely Used Estimator." *Journal of Human Resources*, 32(3): 441–62.
Heckman, James J. 1999. "Instrumental Variables: Response." *Journal of Human Resources*, 34(4): 828–37.
Heckman, James J., and Richard Robb Jr. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, ed. James J. Heckman and Burton Singer, 156–245. Cambridge; New York and Sydney: Cambridge University Press.
Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, 9(2): 85–110.
Heckman, James J., and Sergio Urzua. 2009. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." National Bureau of Economic Research Working Paper 14706.
Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics*, 156(1): 27–37.
Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics*, 88(3): 389–432.
Holden, Constance. 1990. "Head Start Enters Adulthood: After 25 Years We Don't Know Much about How Early Childhood Intervention Programs Work, but Current Research Suggests They Should Be Extended Beyond Early Childhood." *Science*, 247(4949): 1400–1402.
Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81(396): 945–60.
Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics*, 125(1–2): 241–70.
Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115(4): 1239–85.
Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions." In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, 221–33. Berkeley: University of California Press.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–75.
Imbens, Guido W., and Karthik Kalyanaraman. 2008. "Optimal Bandwidth Selection in Regression Discontinuity Designs." Unpublished.
Imbens, Guido W., Gary King, David McKenzie, and Geert Ridder. 2009. "On the Benefits of Stratification in Randomized Experiments." Unpublished.
Imbens, Guido W., and Thomas Lemieux. 2008. "Special Issue Editors' Introduction: The Regression Discontinuity Design—Theory and Applications." *Journal of Econometrics*, 142(2): 611–14.
Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. 2001. "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players." *American Economic Review*, 91(4): 778–94.
Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47(1): 5–86.
Kremer, Michael, and Alaka Holla. 2008. "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." Unpublished.
Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, 114(2): 497–532.
LaLonde, Robert J. 1986. "Evaluating the Econometric

Evaluations of Training Programs with Experimental Data." *American Economic Review*, 76(4): 604–20.
Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review*, 73(1): 31–43.
Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics*, 142(2): 675–97.
Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.
Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review*, 80(2): 319–23.
Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge and London: Harvard University Press.
Manski, Charles F. 1996. "Learning About Treatment Effects from Experiments with Random Assignment of Treatments." *Journal of Human Resources*, 31(4): 709–-33.
Manski, Charles F. 1997. "The Mixing Problem in Programme Evaluation." *Review of Economic Studies*, 64(4): 537–53.
Manski, Charles F. 2003. *Partial Identification of Probabilities Distributions*. New York and Heidelberg: Springer.
Manski, Charles F., Gary D. Sandefur, Sara McLanahan, and Daniel Powers. 1992. "Alternative Estimates of the Effect of Family Structure during Adolescence on High School Graduation." *Journal of the American Statistical Association*, 87(417): 25–37.
Matsudaira, Jordan D. 2008. "Mandatory Summer School and Student Achievement." *Journal of Econometrics*, 142(2): 829–50.
McClellan, Mark, Barbara J. McNeil, and Joseph P. Newhouse. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables." *Journal of the American Medical Association*, 272(11): 859–66.
McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.
Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1): 159–217.
Neyman, Jerzy. 1990. "On the Application. of Probability Theory to Agricultural Experiments. Essays on Principles. Section 9." *Statistical Science*, 5(4): 465–72. (Orig. pub. 1923.)
Ravallion, Martin. 2009. "Should the Randomistas Rule?" *The Economists' Voice*, 6(2).
Reid, Constance. 1982. *Neyman: From Life*. New York: Springer.
Rodrik, Dani. 2008. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" Harvard University John F. Kennedy School of Government Working Paper 08-055.
Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94(4): 1055–84.
Rosenbaum, Paul R. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science*, 2(3): 292–306.
Rosenbaum, Paul R. 1995. *Observational Studies*. New York; Heidelberg and London: Springer.
Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.
Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70(1): 41–55.
Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66(5): 688–701.
Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics*, 6(1): 34–58.
Rubin, Donald B. 1990. "Formal Mode of Statistical Inference for Causal Effects." *Journal of Statistical Planning and Inference*, 25(3): 279–92.
Scheffe, Henry. 1970. "Practical Solutions to the Behrens–Fisher Problem." *Journal of the American Statistical Association*, 65(332): 1501–08.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125(1–2): 305–53.
Tinbergen, Jan. 1997. "Determination and Interpretation of Supply Curves: An Example." In *The Foundations of Econometric Analysis*, ed. David F. Hendry and Mary S. Morgan, 233–45. Cambridge and New York: Cambridge University Press.
Todd, Petra E., and Kenneth Wolpin. 2003. "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico." Penn Institute for Economic Research Working Paper 03-022.
White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–38.
Zelen, M. 1979. "A New Design for Randomized Clinical Trials." *New England Journal of Medicine*, 300(22): 1242–45.

This article has been cited by:

1. Henk Folmer, Olof Johansson-Stenman. 2011. Does Environmental Economics Produce Aeroplanes Without Engines? On the Need for an Environmental Social Science. *Environmental and Resource Economics* **48**:3, 337-361. [CrossRef]
2. Lota D. Tamini. 2011. A nonparametric analysis of the impact of agri-environmental advisory activities on best management practice adoption: A case study of Qu??bec. *Ecological Economics* . [CrossRef]
3. C. B. Barrett, M. R. Carter. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy* **32**:4, 515-548. [CrossRef]
4. Andrew Dillon. 2010. Do Differences in the Scale of Irrigation Projects Generate Different Impacts on Poverty and Production?. *Journal of Agricultural Economics* no-no. [CrossRef]