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

Guido W. Imbens†

April 2009

Abstract

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 towards shoring up the internal validity of estimates, and towards 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 Classification: C14, C21, C52

Keywords: Causality, Instrumental Variables, Local Average Treatment Effects, Regression Discontinuity Designs, Internal Validity, External Validity

*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 grant 0820361.

†Department of Economics, Harvard University, M-24 Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: imbens@harvard.edu, http://www.economics.harvard.edu/faculty/imbens/imbens.html.

1 Introduction

Two recent papers, Deaton (2009; Deaton from hereon), and Heckman and 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 Imbens and 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, page 19). Deaton writes: "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. ... 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 it may, and then proclaim that whatever it illuminates is what we were looking for all along," (Deaton, page 10). He 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, page 14).[2] HU wonder whether these researchers are of the opinion: "that disguising identification problems by a statistical pocedure is preferable to an honest discussion of the limits of the data?" (HU, page 19).

The fact that two such distinguished economists so forcefully 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 this is not the case. Much progress has in fact been made in empirical practice, and empirical work is much more credible as a result of the natural experiments revolution started by Card, Angrist, Krueger, and others in the late eighties. Starting in the late eighties their work, and more recently that by development economists such as Banerjee, Duflo, and Kremer arguing in favor of randomized experiments, has had a profound influence on empirical work. By emphasizing internal validity and study design, this literature showed the importance of looking

[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)

[2] 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, page 2).

for clear and exogenous sources of variation in potential causes. In contrast to what Deaton and HU suggest, this issue is distinct and separate from the choice of the models and estimation methods used. In fact, recently there has been much interesting work exploring the benefits of randomization for identification, estimation and assesment of structural models. For an early example see Hausman and Wise (1979) who estimate a model for attrition with data from randomized income maintenance experiment, and for recent examples see Card and Hyslop (2005) who estimate a structural model of welfare participation using experimental data from Canada, Todd and Wolpin (2003), who analyze data from Mexico's Progressa program, Imbens, Rubin and Sacerdote (2001) who estimate labor supply models exploiting random variation in unearned earnings using data from lottery winners, Duflo, Hanna, and Ryan (2007) who look at the effect of monitoring and financial incentives on teacher's absences, and Athey, Levin and Seira (2004) who use randomized assignment of auction formats to estimate structural models of bidding behavior. There is much room for such work where experimental variation is used to improve the identification 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 minor squabbles about the relative merits of structural work versus work less directly connected to economic theory, or in discussions of second-order technicalities such as adjustments for heteroskedasticity in the calculation of standard errors and the Behrens-Fisher problem (e.g., Deaton, page 33).[3]

In my view, it is helpful to separate the discussion into two parts. The first part concerns the questions of interest, and the second the methods conditional on the question. In my opinion the main concern with the current trend towards credible causal inference in general, and towards randomized experiments in particular, is that it may lead researchers to avoid questions where randomization is difficult, or even conceptually impossible. There are many such questions, and many of them are of great importance. Questions concerning the causal effects of macro-economic policies can rarely be settled by randomized experiments. 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

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

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, Rodrik (2008). There is clearly much room for non-experimental 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 towards general acceptance of the causal interpretation of the smoking and lung cancer correlation (and Fisher's long-time scepticism about the causality 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 David Card calls design-based strategies, have many advantages. Specifically, conditional on the question of interest being one for which randomized experiment is feasible, randomized experiments are superior to all other designs in terms of statistical reliability. Where as Deaton sees no special role for experiments, Freedman, hailed by Deaton himself as "one of [the world's] greatest statisticians" (Deaton, title page, acknowledgement) is unambiguous in his opening sentence, "Experiments offer more reliable evidence on causation than observational studies," (Freedman, 2006, abstract) 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. However, no other design will have the credibility that a randomized experiment would have. Suppose we are interested in 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 for some recent discussions Holden (1990) and Manski (1997)), the National Supported Work

Demonstration experiments on labor market programs (e.g., 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 that up 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 endogeneity.

In these comments I will make five points, from the perspective of an econometrician who is interested 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 partially takes its motivation 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 Fraker and Maynard (1987). Third, I will argue 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 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. An important insight is that in settings with heterogenous effects, instrumental variables strategies do not identify the average effect of the treatment (e.g., Heckman, 1990). However, as shown by IA instrumental variables methods do identify the average treatment effect for a well defined subpopulation, the average effect for what IA call the compliers. Although in many cases the local average treatment effects, 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 restrictions that underly 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 invariant, or structural, than the latter, that is, are more likely to generalize to other settings.

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 overall population with little credibility. This is consistent with the biomedical literature. Although the primacy of internal validity over external validity has been criticized in that literature, there is little support for moving towards a system where studies with low internal validity receive much weight in policy decisions. External validity is generally a bigger problem in economics than in biomedical settings, with substantial variation in both preferences and constraints between individuals, as well as variation over time. Understanding variation in treatment effects is therefore of great importance in these settings, and it has received a fair amount of attention in the experimental literature (e.g., Banerjee and Duflo, 2008).

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 reduced-form, atheoretical, or statistical (as opposed to economic). These are not terms commonly used in this literature itself. They are also at odds with their historical use.[4] In the classical simultaneous equations setting, the 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), not to equations estimated by instrumental variables methods. The almost complete lack of instrumental variables methods in the statistical literature makes that

[4]In an even more remarkable attempt to shape the debate by changing terminology, Deaton proposes to redefine exogeneity in a way that allows for the possibility that a randomly generated number is not exogenous with respect to economic behavior.

label also inappropriate 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 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 Lee and Lemieux (2009) 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 (e.g., Duflo, Glennerster, and Kremer, 2008). 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., Bertrand and Mullainathan (2004), Duflo and Saez (2003), and many others.

Often the focus 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. Even if simple average effects of these interventions are not directly the answering questions about plausible economic policies, they are often closely related to the effects of 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 the assignment to a particular regime was 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 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%. En-

tering in the California policymakers' decision is the a 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 is 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 Lavy, 1999). Third, there are estimates exploiting natural variation in class size arising from natural variation in cohort size, using data from Connecticut in 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 policy maker? 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 towards 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 to the studies relative to the target program, with older studies receiving less weight, (ii) differences between the study population and the California target population, including the targed grade levels in each study, (iii) differences between the study outcomes and the goals of the California programs. In terms of these criteria the Connecticut study would do best. In terms of internal validity, that is, of the estimate having a credible causal interpretation, the experimental Tennessee study and, next, the Israeli study would 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 policy makers at this moment very much. If we did an observational study in California,

however, I would still put some weight on the Connecticut, Tenessee 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 policy makers even if none of the individual studies directly answers the policy question of interest. It is of course unrealistic to expect that the California policy makers 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, 2008). 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. 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, Duflo, Cole, and Linden (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 that 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 for another example of such a meta analysis Card, Kluve, and Weber (2009), and for additional discussion Section 6.

3 Lalonde (1986): The Failure of Non-experimental 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 modelling. HU vaguely reference the "practical difficulty in identifying, and precisely estimating the full array of structural parameters" (HU, page 2), but mention only an unreferenced 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, page 2, footnote 6). I think the origins behind this flight are not quite as obscure 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." In this 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 non-experimental data. He concluded that they could not do so systematically. Lalonde's evidence, and subsequent confirmations of his 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 seems clear 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 policy makers 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 Dehejia and Wahba (1999), although this is still disputed, e.g., Smith and Todd (2005). See Imbens and Wooldridge (2009) for a recent review. Matching methods, however, hardly meet Deaton's criteria for "analysis of models inspired by economic theory" (Deaton, page 2). Until there are more succesful 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.

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 experiments has no special priority. ... Randomized experiments cannot automatically trump other evidence, they do not occupy any special place in some hierarchy of evidence," (Deaton, page 4). 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.[5]

Formally, as shown originally by Fisher (1925), randomization allows the researcher to precisely quantify the uncertainty associated with the evidence for an effect of a treatment. 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, assumptions on the sampling process, or assumptions on interactions between units, 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 subtle extensions to more interesting questions. We can establish the presence of a causal effect through the calculation of p-values, but we cannot estimate the average effect without some additional assumptions. Unless we rule out interactions, the average effect depends on assignments to other individuals

[5]See the earlier quote by Freedman.

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 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 his earlier comments there is no initial superiority to undermine. It is obviously true that violations of assignment protocols, missing data, and other practical problems, 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. Moreoever, 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 $2N$ units, N randomly assigned to the treatment group, and the remaining N assigned to the control group. Suppose we wish to test the sharp Fisher null hypothesis that there is no causal effect whatsoever of a job search assistance program on employment status. For such an experiment we can calculate the exact p-values using Fisher's methods. Now suppose that there is noncompliance. Some individuals assigned to the program, did not participate in the program, and some assigned to the control group, did in fact participate in the program. Let $Y_i^* \in \{0,1\}$ be the outcome we would have observed for individual i, had this individual been exposed to the treatment assigned to her. Let C_i be an indicator for compliance with the treatment received, and let W_i be the treatment assigned. The complete data p-value p^{comp} can be written as a function the complete data, $p^{\text{comp}} = p(\mathbf{Y}^*, \mathbf{W})$, where $\mathbf{Y}^*$ and $\mathbf{W}$ are the $2N$ vectors with typical element Y_i^* and W_i respectively. The problem is that we do not observe Y_i^* if $C_i = 0$ (the individual does not comply with the treatment assigned). However, even in that case we know that $Y_i^* \in \{0,1\}$. Hence we can derive, in the spirit of the work by Manski (1990,

1994, 2003), the range of p-values consistent with the observed data, without making any assumptions whatsoever about the nature of the noncompliance. Depending on the data, we may therefore be able to conclude, in spite of the noncompliance, that we can be confident that the treatment did have some effect. The point is that in settings with limited noncompliance we can still make precise statements of the type validated by randomized experiments, with no 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 pratical complications.

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 policy makers as the mean treatment effect." Deaton is correct in stating that experiments are not informative about the median treatment effect. As a side issue, this raises the question, of course, how any study can be, other than by making untestable assumptions, but let me ignore that question. The more important issue is the second claim in the Deaton quote. In many cases average effects on (functions of) outcomes are indeed what is of interest to policy makers, *not* quantiles of differences in potential outcomes. The key insight is that 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 distribution of outcomes induced by the first program, and compare that to the welfare induced by the second program. 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, page 714). (Here $P[Y(w)]$ denotes the distribution of $Y(w)$.) The 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 $Y_i(1) - Y_i(0)$.

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.[6] Deaton raises three issues. The first concerns inference, or 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, note, however, that although Deaton raises these issues in the context of experimental evaluations, there is nothing specific to randomized experiments that makes them more vulnerable to these issues than observational studies. Moreoever, in my view these three are second order issues. That is, second order relative to the first order issues of selection and endogeneity in observational evaluation studies that have long been highlighted by Heckman (e.g., Heckman, 1978; Heckman and Robb, 1985).

First, 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. Using the standard ols variance based on homoskedasticity leads to confidence intervals that are not necessarily justified even in large samples. This is correct, and in practice it is certainly recommended to use the robust variance here. Moroever, the standard error issue that is often the biggest 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 include covariates in regression estimates of average treatment effect. In randomized experiments this is not strictly necessary. Because the covariates are in expectation uncorrelated with the treatment indicator, the standard omitted variable bias argument implies that their omission or inclusion does not introduce any asymptotic bias. In finite samples including covariates can introduce some bias, because the finite sample correlation between the treatment indicator and the covariates need not equal zero, even if the population correlation does. On the other hand, including covariates can substantially improve the precision if these covariates are good predictors of the outcomes given or without the treatment. In finite samples there is therefore a tradeoff between some finite sample bias, and large sample

[6] In fact one would always, even in small samples be at least as well off by stratification on these covariates, e.g., Imbens, King, McKenzie, and Ridder (2009).

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.[7]

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 researchers uses pre-testing, or estimates multiple versions of a statistical model there should be concern that the final confidence intervals no longer have the nominal coverage rate. However, I think that this is again a second order issue. 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, 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 their own treatment status. In other words, conditional on the question, the methodological case for randomized experiments seems unassailable, 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 given 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.

[7] 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 even fewer observational settings for which we have exact finite sample results.

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. 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.[8] In doing so, this literature has made many connections to the statistics and psychology literature on observational studies. Rather than leading to "unnecessarily rhetorical barriers between disciplines" (Deaton, page 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, 1973, 1990; Rosenbaum and Rubin, 1983), labeled the Rubin Causal Model by Holland (1986), which formulates causal questions as comparisons of unit-level potential outcomes.[9] Although this 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, Haavelmo (1943). 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 variables published in the mainstream medical literature, although still written by economists, McClellan and 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., Zelen, 1979), but no links to the econometrics literature had been made. Furthermore economists significantly generalized applicability and understanding of regression

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

[9] Compare for example, the set up in Heckman and Robb (1985) with that in Heckman (1990).

discontinuity designs (Hahn, Todd, and VanderKlaauw, 2001), which originated in the psychology literature. See Shadish, Campbell and Cook (2000), and Cook (2008) for a historical perspective. Within economics, however, the results in IA and Hahn, Todd, and VanderKlaauw (2001) are unusual. 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, at the outset of an analysis, what is the object of interest. Let me use Angrist's (1989) 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 non-veterans are not credible as estimates of average causal effects because of 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.[10] This is a substantive assumption that may well be violated. See Angrist (1990) and Angrist, Imbens, and Rubin (1996) for discussions of potential violations. 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.[11] In this setting monotonicity seems a very reasonable assumption. See Angrist, Imbens, and Rubin for discussions of 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, we can identify the average effect on the subpopulation of what Angrist, Imbens and Rubin (1996) call compliers. 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.

[10]Deaton actually calls this second assumption "exogeneity", in an unnecessary and confusing change from conventional terminology

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

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 the 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 us 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, 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 $[-1/4, 1/4]$. This can be consistent with a substantial negative average effect for compliers, lowering survival rates by 1/8, or with a substantial positive average effect for compliers, raising survival rates by 1/8.[12] In both examples there need not be any statistical evidence that the effect differs for compliers and nevertakers. One would think that in the first case a decision maker would be considerably less likely to implement universal adoption of the treatment than in the second, and so reporting only

[12]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, $p_{00} = 3/8$, $p_{10} = 1/4$, and $p_{11} = 1/4$. In the second example $\tilde{p}_{00} = 1/8$, $\tilde{p}_{10} = 1/4$, and $\tilde{p}_{11} = 1/4$. Now in both cases the sharp bounds on the average treatment effect are $[-1/4, 1/4]$, in the first example $\tau_{\text{late}} = -1/8$, and in the second example $\tilde{\tau}_{\text{late}} = 1/8$.

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 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 (earnings), depending on veteran status:

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

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 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 $\eta_i < -\pi_0 - \pi_1$ (nevertakers in the AIR terminology) would not serve even if drafted, and individuals with $-\pi_0 - \pi \leq \eta_i < -\pi_0$ (compliers) would serve if drafted, but not as volunteers, and

individuals with $-\pi_0 \leq \eta_i$ (alwaystakers) would always serve.[13]

Although not widely used anymore, this type of model was very popular in the eighties, 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 assumption 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 identyfing 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 $(\varepsilon_i, \eta_i, \nu_i)$), we can still identify the parameters of the model, including β, that is, the average effect of veteran status. See for example, Björklund and 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" (Chamberlain, 1986; HU). In the case with a binary instrument, this assumption is easy to check. In the Angrist study, the probability of serving in the military for the draft eligible and non-eligible is far from zero and one, and so nonparametric identification fails. The contribution of the LATE literature 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. In the structural model above, compliers are the individuals with $\pi_0 - \pi_1 \leq \eta_i < \pi_0$. Think again

[13] There are also arguments why the instrument need not be valid. For example, individuals may avoid the draft by enrolling in additional education to receive educational deferments. See Angrist, Imbens and Rubin (1996) for more discussion.

of the case where the nevertakers 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, page 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 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, learn 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 separate.

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 Lee's (2008) example of the effect of incumbency on election outcomes. Lee uses comparisons of congressional districts where the previous election was barely won by a Democrat with districts where the previous election was barely wone by a Republican. This leads to estimates of the effect of incumbency that have a high degree of internal validity, but that only apply to districts with close elections. These may well be very different from districts that are heavily leaning to one party. There is little reason to believe that districts with close elections are the only ones of interest, but in the absence of credible models for extrapolation, this

is again all we can do.

Fuzzy regression discontinuity designs rank even lower in terms of external validity. As pointed out by Hahn, Todd, and VanderKlaauw (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 choose to report estimates for them.

6 Internal versus External Validity

Much of the debate ultimately centers on the weight researchers put on external validity versus internal validity of estimators. There is no disagreement that both are important. See Banerjee and Duflo (2008) 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 soley 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 treatments with a subsequent reversal of that decision, after the treatment was found to have adverse effects on populations that were

underrepresented in the original study populations. Part of this is unavoidable. First, randomized experiments can only be conducted on volunteers, and there is no systematic method for ensuring that the population of volunteers is representative of the population of interest. Second, after a succesful randomized experiment, the target population may well change. If a treatment is in a trial very succesful for moderately sick patients, it may well be used for sicker patients that were not part of the original study. Doing a second experiment is not always an option, and is often not ethical if there are demonstrable and sizable effects on a closely related population. 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 generalize to other populations. A vaccine for a particular strain of HIV that prevents infection in the US 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 US given the differences in educational institutions and practices.

It may be helpful to put some more structure on this problem.[14] 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. 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, 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 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

[14] This discussion is partly based on conversations with Abhijit Banerjee and Sendhil Mullainathan.

to all states. Given this model, the difference in expected demand in state s if the price is fixed at p_1 versus p_0 is

$$\theta_s = \mathbb{E}[D|S=s, P=p_1] - \mathbb{E}[D|S=s, P=p_0] = \beta_1 + \beta_2 \cdot \mathbb{E}[I|S=s].$$

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

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

Futhermore, suppose that our interest is solely in the difference in average outcomes in California,

$$\theta_{\text{ca}} = g(X_{\text{ca}}, \beta).$$

Now consider the case where 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 $\theta_{\text{tn}} = g(X_{\text{tn}}, \beta)$.

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 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 $\hat{\beta}_{\text{ct}}$, 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,

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

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

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

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 $\hat{\theta}^{\rm exp}_{\rm ca}$. If one believed the structural model was close to correctly specified, one would prefer $\hat{\theta}^{\rm struct}_{\rm ca}$. 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 $\beta_{\rm ct}$ would be equal to β, and thus $g(X_{\rm ca}, \beta_{\rm ct})$ would be equal to $\theta_{\rm ca}$. 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.

The answer to the first question may also differ if the experiment in Tennesee 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 $\theta_{\rm ca}$ 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 Tennesee, $\theta_{\rm tn}$, with the structural one based on Connecticut estimates, $\hat{\theta}^{\rm struct}_{\rm tn} = g(X_{\rm tn}, \beta_{\rm ct})$. The comparison of $\theta_{\rm tn}$ and $\hat{\theta}^{\rm struct}_{\rm tn}$ 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 $\hat{\theta}^{\rm struct}_{\rm ca}$. This test does not reflect in any way on the experimental estimate $\hat{\theta}^{\rm exp}_{\rm ca}$.

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 $\mathbb{E}[\theta_s|X_s] = g(X_s)$. 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), Hotz, Imbens and Mortimer (2005), Kremer and Holla (2008) and Duflo nand Chattopadhyay (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.

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 towards 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 towards improving the credibility of empirical work.

References

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J., G. Imbens and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J., and V. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, Vol. CXIV, 1243.

Athey, S., J. Levin, and E. Seira, (2004), "Comparing Open and Sealed Bid Auctions: Theory and Evidence from Timber Auctions,"Unpublished Paper.

Banerjee, A., (2007), *Making Aid Work*, Cambridge, MIT Press.

Banerjee, A., and E. Duflo, (2008), "The Experimental Approach to Development Economics,"unpublished manuscript, department of economics, MIT.

Banerjee, A., and R. He, (2008), Making aid work, in William R. Easterly, ed., *Reinventing Foreign Aid*, Cambridge, MA. MIT Press, pp. 4792.

Banerjee, A., E. Duflo, S. Cole, and L. Linden, (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics* Vol. Vol. 122(3): 1235-1264.

Bertrand, M., and S. Mullainathan, (2004), "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, Vol. 94(4):991-1,013.

Björklund, A. and R. Moffitt, (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models," *Review of Economics and Statistics*, Vol. LXIX, 42–49.

Card, D., (1990), "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43, 245-257.

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

Card, D., J. Kluve, and A. Weber, (2009), "Active Labor Market Policy Evaluations: A Meta-Analysis,"Working Paper 4002, Institute for the Study of Labor (IZA).

Card, D., and A. Krueger, (1994): "Minimum Wages and Employment: A Case Study of the Fast-food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84 (4), 772-784.

Chattopadhyay, R., and E. Duflo, (2004) "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica*, Volume 72(5):1409-1443.

Chamberlain, G. (1986), "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, Vol 32(2), 189-218.

Cook, T., (2008), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, *Journal of Econometrics*. Vol 142(2):636-654

Deaton, A., (2009), "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development,"NBER Working Paper #14690.

Dehejia, R., and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 1053-1062.

Duflo, E. (2004), Scaling up and evaluation, Annual World Bank Conference on Development Economics 2004, Washington, DC. The World Bank.

Duflo, E., R. Glennerster, and M. Kremer, (2008), "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, (T. P. Schultz and J. Strauss eds.), 3895-3962.

Duflo, E., R. Hanna, and S. Ryan (2007), "Monitoring Works: Getting Teachers to Come to School," unpublished manuscript, department of economics, MIT.

Duflo, E., and E. Saez (2003): "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 118(3):815-842, 118(3).

Fisher, R. A., (1925), *The Design of Experiments*, 1st ed, Oliver and Boyd, London.

Fraker, T., and R. Maynard, (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources*, Vol. 22, No. 2, p 194–227.

Freedman, D., (2006), "Statistical Models for Causation: What Inferential Leverage Do They Provide," *Evaluation Review*, 691-713.

Freedman, D., (2008), "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics*, 180-193.

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica* 11, 1-12.

Hahn, J., P. Todd, and W. van der Klaauw, (2000), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1): 201-209.

Hausman, J., (1981), "Labor Supply," in Aaron and Pechman (eds.) *How Taxes Affect Economic Behavior*, The Brooking Institution, 27-72.

Hausman, J., and D. Wise, (1979), "Attritin Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, Vol. 47(2): 455-473.

Heckman, J., (1978), "Dummy Endogenous Variables in a Simultaneous Equations System", *Econometrica*, Vol. 46, 931–61.

Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review* 80, 313-318.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., and J. Smith, (1995), "Assessing the Case for Social Experiments," *Journal of Economic Perspective*, 9(2), 85-115.

Heckman, J., and S. Urzua., (2009), "Comparing IV With Structural Models: What Simple IV Can and Cannot Identify," NBER Working Paper, # 14706.

Holden, C., (1990), "Headstart Enters Adulthood," *Science*, 247: 1400-1402.

Holland, P., (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-970.

Hotz, V. J., G. Imbens, and J. Mortimer, (2005), "Predicting the Efficacy of Future Training Programs Using Past Experiences", *Journal of Econometrics*, vol 125(1-2):241-270.

Hoxby, C. (2000), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, Vol. 115(4):1239-1285.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., G. King, D. McKenzie, and G. Ridder, (2009), "On the Benefits of Stratification in Randomized Experiments," unpublished manuscript, Department of Economics, Harvard University.

Imbens, G., D. Rubin, and B. Sacerdote, (2001), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91, 778-794.

Imbens, G., and T. Lemieux (2008) "Special issue editors' introduction: The regression discontinuity design - Theory and applications," *Journal of Econometrics*, Vol 142(2), 611-644.

Imbens, G., and J. Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation," forthcoming, *Journal of Economic Literature*.

Kremer, M., and A. Holla, (2008) "Pricing and Access: Lessons from Randomized Evaluations in Education and Health,", unpublished manuscript, department of economics, Harvard University.

Krueger, A. (1999) "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, Vol 114(2), 497-532.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.

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

Lee, D. and T. Lemieux, (2009), "Regression Discontinuity Designs in Economics ,"NBER Working Paper # 14723.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C., (1997), "The Mixing Problem in Programme Evaluation," *Review of Economic Studies*, 64: 537:553.

Manski, C., (1996), "Learning about Treatment Effects from Experiments with Random Assignment of Treatments," *The Journal of Human Resources*, 31(4): 709-73.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

McClellan, M., and J. P. Newhouse, (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality", *Journal of the American Medical Association*, Vol 272, No 11, 859–866.

Neyman, J., (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,"translated in *Statistical Science,* (with discussion), Vol 5, No 4, 465480, 1990.

Rodrik, D., (2008), "The New Development Economics: We Shall Experiment, But How Shall We Learn?,"Unpublished Manuscript, Kennedy School, Harvard University.

Rosenbaum, P., (1995), *Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. , (1978), "Bayesian inference for causal effects: The Role of Randomization", Annals of Statistics, 6:34–58.

Rubin, D. B., (1990), "Formal Modes of Statistical Inference for Causal Effects," Journal of Statistical Planning and Inference, 25, 279-292.

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton and Mifflin, Boston.

Smith, J. A. and P. E. Todd, (2005), "Does Matching Address LaLonde's Critique of Nonexperimental Estimators," *Journal of Econometrics*, 125(1-2), 305-353.

Todd, P., and K. 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

Zelen, M., (1979), "A New Design for Randomized Clinical Trials", *New England Journal of Medicine*, 300, 1242–1245.