Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Rejoinder

Joshua D. Angrist , Guido W. Imbens & Donald B. Rubin

To cite this article: Joshua D. Angrist , Guido W. Imbens & Donald B. Rubin (1996) Rejoinder, Journal of the American Statistical Association, 91:434, 468-472, DOI: 10.1080/01621459.1996.10476907

To link to this article: https://doi.org/10.1080/01621459.1996.10476907

Holland, P. (1988), "Causal Inference, Path Analysis, and Recursive Structural Equation Models," in *Sociological Methodology*, Washington, DC: American Sociological Association.

Hollander, M., and Wolfe, D. (1973), *Nonparametric Statistical Methods*, New York: John Wiley.

Kempthorne, O. (1952), *Design and Analysis of Experiments*, New York: John Wiley.

Lehmann, E. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden Day.

Maritz, J. S. (1995), *Distribution-Free Statistical Methods*, London: Chapman and Hall.

Rosenbaum, P. R. (1993), "Hodges–Lehmann Point Estimates of Treatment Effect in Observational Studies," *Journal of the American Statistical Association*, 88, 1250–1253.

——— (1995), *Observational Studies*, New York: Springer-Verlag.

Rejoinder

Joshua D. ANGRIST, Guido W. IMBENS, and Donald B. RUBIN

We thank Heckman, Greenland and Robins, Moffitt, and Rosenbaum for their stimulating comments on our paper. After making two general remarks, we address specific points in each comment.

Both Heckman and Greenland and Robins stress that LATE is the average causal effect for a subpopulation that cannot be identified in the sense that we cannot label all individual units in the population as compliers or noncompliers. Greenland and Robins suggest that attention should focus on the population average treatment effect, whereas Heckman is more interested in the average effect for those who receive treatment, also the estimand of interest in Peters (1941), Belson, (1951), Cochran (1969), and Rubin (1973a,b, 1977). For policy purposes, one may indeed be interested in averages for the entire population, or for specific subpopulations other than compliers. Within the context of a particular study with a specific instrument, however, the data are not directly informative about average effects for subpopulations other than compliers. A key insight from our work is that compliers are the *only* group with members observed taking the treatment and members observed not taking the treatment. Always-takers are always observed taking the treatment, so the data simply cannot be informative about average treatment effects for this group, and similarly for never-takers. In the same vein, a clinical trial restricted to young men is not informative about treatment effects for adult women. Yet Heckman and Greenland and Robins appear to criticize us precisely because we limit our discussion of causal effects to the only subpopulation about which the data are directly informative.

Following a core analysis focused on the directly estimable effect, one may wish to extend the conclusions to broader groups. Such extensions are routine in the interpretation of clinical trials, which are seldom based on representative samples of the overall target population. Our approach makes it clear, however, that in instrumental variables (IV) contexts, extensions to groups other than compliers can only be extrapolations.

The second issue raised by multiple discussants is the propriety of our example. Clearly, an example with a binary randomized instrument is not representative of economic applications of IV techniques where candidate instruments are rarely based on actual randomization. A major reason for using this example was to stress that randomization alone does not make a candidate instrument a valid one because randomization does not make the exclusion restriction more plausible. The fact that economists do not always make a clear distinction between ignorability and exclusion restrictions is evidenced by Moffitt's incorrect comment that randomization makes the draft lottery "by *necessity* an obvious and convincing instrument" (italics ours) for the effect of the military service. In fact, one contribution of our approach is to provide a framework that clearly separates ignorability and exclusion assumptions. Both statisticians and economists should find this separation useful and clarifying.

HECKMAN

Heckman begins by arguing that the RCM is a version of the widely used econometric switching regression model. We view the term Rubin causal model (coined by Holland [1986] for work by Rubin [1974, 1978]) as referring to a model for causal inference where causal effects are defined explicitly by comparing potential outcomes. This comparison can be in the context of a randomized experiment or an observational study. Any element of the set of the potential outcomes *could* have been observed by manipulation of the treatment of interest, even though ex-post only one of them is actually observed. Moreover, the RCM defines the assignment mechanism, which determines which potential outcomes are observed, as the conditional probability of each possible treatment assignment given the potential outcomes and possibly other variables. In contrast, the switching regression model as exposited by Quandt (1958, 1972) is a time series model where the first part of the sample comes from one regression model and the second part from a separate regression model with an unknown switching point.

A second example mentioned by Heckman is Roy (1951), who studied the distribution of observed incomes in a world

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Applications and Case Studies

where individuals always choose the occupation with the highest income. Neither Roy (1951) nor Quandt (1958, 1972) discussed causal effects. What makes the Roy model and the switching regression model technically closer to the RCM than many models used in econometric evaluations studies (e.g., many of the models in Heckman and Robb 1985) is their explicit focus on potential outcomes as distinct from observed outcomes. Only recently has the RCM potential outcome framework been adopted in economic models for causal effects (e.g., Maddala 1983, Bjorklund and Moffitt 1987, Heckman 1990, and Manski 1990). Once potential outcomes have been introduced, one can indeed define disturbances as deviations of these potential outcomes from their population expectations, as Heckman does in his comment. Our remarks regarding the difficulty in interpretating these disturbances (e.g., the Holland 1988 quote given in our article), refer to papers where the disturbances are used but are not defined in terms of potential outcomes.

In statistics (e.g., Fisher 1918, Neyman 1923, and other early references provided in Rubin 1990), as well as in economics, there are studies that contain elements of the RCM. Two early economic examples that we find more relevant than either the Roy or Quandt articles cited by Heckman are Tinbergen (1930) and Haavelmo (1944), both founders of modern econometrics. Tinbergen wrote: "Let π be any imaginable price; and call total demand at this price $n(\pi)$, and total supply $a(\pi)$. Then the actual price p is determined by the equation $a(p) = n(p)$, so that the actual quantity demanded, or supplied, obeys the condition $u = a(p) = n(p)$.... The problem of determining demand and supply curves ... may generally be put as follows: Given p and u as functions of time, what are the functions $n(\pi)$ and $a(\pi)$?" (Tinbergen 1930, translated in Hendry and Morgan 1995, p. 233). This very clearly describes the potential outcomes and the specific assignment mechanism corresponding to market clearing, although there is no statistical model in Tinbergen's discussion. Similarly, Haavelmo wrote: "When we set up a system of theoretical relationships and use economic names for the otherwise purely theoretical variables involved, we have in mind some actual *experiment,* or some *design of an experiment,* which we could at least imagine arranging, in order to measure those quantities in real economic life that we thank might obey the laws imposed on their theoretical namesakes." (Haavelmo 1994, p. 6, italics in original). Although more ambiguous than the Tinbergen quote, this certainly suggests that Haavelmo viewed laws or structural equations in terms of potential outcomes that could have been observed by "arranging" an experiment.

In his Section 2 Heckman provides an alternative set of assumptions for identification of the average effect on the treated, arguing that these assumptions are more transparent and have more behavioral content than our assumptions. As discussed in the Introduction to our reply, our focus on the complier average causal effect is not incidental, nor do we view compliers as the only interesting group. Rather, we focus on the average causal effect for compliers because this is the only directly estimable causal effect of the treatment. The only way to get average effects for always-takers and never-takers is to assume that their average treatment effects can be deduced from those for compliers, and this is exactly what Heckman has done in his assumptions without being explicit about it.

To formalize this argument, let us rewrite Heckman's assumption (A-2′) in our potential outcome notation:

$$E[Y_i(1) - Y_i(0)|Z_i, D_i(Z_i) = 1] = E[Y_i(1) - Y_i(0)|D_i(Z_i) = 1],$$

where we drop the predictor or attribute X from the discussion because all substantive points can be made in the simple case without predictor variables. Consider the impact of this assumption given random assignment of Z, the exclusion restriction, and the monotonicity assumption (implied by most econometric models). Simple manipulation shows that Heckman's assumption (A-2′) implies that

$$E[Y_i(1) - Y_i(0)|D_i(0) = D_i(1) = 1] = E[Y_i(1) - Y_i(0)|D_i(0) = 0, D_i(1) = 1].$$

In words, Heckman's assumption (A-2′) amounts to assuming that the effect for always-takers is the same as that for compliers. Given this assumption, Heckman claims that he can identify a more interesting parameter: the average effect for those who receive the treatment. But because those who receive the treatment are a mixture of always-takers and compliers, Heckman's assumptions simply assume the answer. In the draft lottery example, Heckman's assumption implies that the average effect of military service for volunteers is the same as that for draftees, an assumption that we carefully avoided in Angrist (1990) and in our work.

We also view Heckman's assumption (A-2′) as lacking in scientific (economic) content. Our assumptions restrict outcomes at the unit level given different assignments, so that—like Fisher (1918), Neyman (1923), Tinbergen (1930) and Haavelmo (1944)—we compare *for a specific unit* the outcomes that would be observed given different environments. Thus our assumptions can be immediately interpretated as comparisons of outcomes in behavioral models of utility maximizing behavior given different sets of constraints. In contrast, Heckman's key assumption (A-2′) compares *average* outcomes for *different* groups of individuals. He provides no examples where this assumption is plausible or can be related to the economic behavior of agents.

Heckman also takes issue with our ignorability assumption, arguing that mean independence is weaker than full independence. The second assumption obviously implies the first. However, as we argue elsewhere in more detail (Imbens and Rubin 1994), this distinction is not meaningful in practice. If mean independence holds but full independence does not hold, then Z would be a valid instrument for the effect of D on Y but not for a transformation of Y such as $\log(y)$. It would inevitably tie the validity of the instrument to the specific form of the regression function, and return to the functional–form–dependent approach to instrumental variables that we avoid.

A secondary point in Heckman's Section 3 concerns his connection between ignorability and "Granger noncausality." Holland (1986) and Granger (1986, in his comment on the Holland paper) discussed Granger causality in terms of the potential outcomes framework. Heckman's view of Granger causality appears to differ from those of either Granger or Holland.

Heckman's main point in his Section 3 concerns the example and the appropriateness of IV methods in general. These comments are in marked contrast with his earlier views, as expressed in Heckman and Robb (1985): "The instrumental variables estimator is the least demanding in the a priori conditions that must be satisfied for its use .... It is important to notice how weak these conditions are" (p. 185). In contrast to this earlier view, Heckman's current view supports our position that instrumental variables assumptions are strong. Our concern with making such strong assumptions in practice motivates these sensitivity analyses and related discussion in Section 6, where we present possible reasons why the IV assumptions need not be satisfied in our example. Heckman's specific argument is merely another possible reason to believe the exclusion restriction may be violated. Although we have discussed possible violations of the key assumptions at length, we still view the draft lottery example as one of the most convincing examples of IV methods in the literature. In this case the exclusion restriction certainly appears more reasonable than the alternative assumption of ignorable treatment, which would imply that valid causal inferences could be drawn from direct comparisons of veterans and nonveterans.

We also find the draft lottery example more convincing than the application in Robinson (1989), cited by Heckman as an example where "application of IV can be justified in the context of heterogeneous treatments." We view the Robinson study as an example of an IV application where the critical assumptions are formulated in a way that makes it almost impossible to judge their plausibility. For example, Robinson defines endogeneity as a restriction on the covariance of a disturbance and a function of three disturbances. In contrast, our formulation casts the ignorability assumption in terms of independence of the candidate instrument and potential outcomes, and the exclusion restriction in terms of the effect of specific manipulations on observed outcomes. Most importantly, despite their crucial role, the instruments in the Robinson study are never clearly defined and appear to be solely nonlinear functions of the predictor variables.

Heckman's current pessimistic view of IV methods can also be contrasted with the development of his views on a class of experimental evaluation designs with randomized eligibility. In these designs, units are randomly assigned an instrument Z_i with $Z_i = 1$ implying that unit i is eligible for a particular treatment and $Z_i = 0$ implying that unit i is not eligible to receive treatment. Formally, this is a special case of our model with $D_i(0) = 0$ (no defiers or always-takers), and hence monotonicity is automatically satisfied. An alternative interpretation of this example is as a clinical trial with one-sided noncompliance. The exclusion restriction requires that for those who do not take the treatment if eligible, there is no effect of the assignment. Although this is a strong assumption, which need not be satisfied in all cases with randomized eligibility, it can be plausible in many cases, especially in double-blind trials. Given the exclusion restriction, and with the other assumptions satisfied by definition, our IV approach can be used to estimate the average effect for compliers; that is, those who take the treatment when eligible. Because all those observed to take the treatment must be eligible, the IV estimand, LATE (the average effect for compliers) is equal to the "average effect on the treated." Zelen (1979) and Bloom (1984) discussed evaluations based on such designs, and Angrist and Imbens (1991) and Imbens and Angrist (1994) pointed out the connection with instrumental variables.

Heckman's (1991) original discussion of such randomized eligibility designs ignored IV methods and stated only that "a simple mean difference comparison between treated and untreated persons is *less* biased for $E[\Delta|D = 1]$ than would be produced from a mean difference comparison between treated and untreated samples without randomized eligibility. In general, the simple mean difference estimator will still be biased" (p. 27, emphasis in original). More recently, Heckman (1995, pp. 9–10) acknowledged that IV methods can be used to estimate interesting average treatment effects in this context. Specifically, he writes that "this type of randomization [of eligibility] can be placed in an instrumental variables framework .... Note that this type of randomization identifies $E[\Delta|D = 1, X]$." In this context Heckman's estimand $E[\Delta|D = 1, X]$ is actually the same as the local average treatment effect for units with covariate values X.

ROBINS AND GREENLAND

Robins and Greenland offer several alternative analytic strategies, focusing on estimation of bounds for the population average treatment effect. Our approach can also be used to generate bounds on the population average treatment effect in a straightforward fashion. Given monotonicity, there are three groups: compliers, always-takers, and never-takers. Given random assignment, we know in large samples the population fraction of the three types, and moreover, given the exclusion restriction, we know the average treatment effect $Y_i(1) - Y_i(0)$ for compliers, the average of $Y_i(1)$ for always-takers, and the average of $Y_i(0)$ for never-takers. Without further assumptions, the two unknown components in the population average treatment effect are the average of $Y_i(0)$ for always-takers and the average of $Y_i(1)$ for never-takers. The data contain no direct information about these two quantities. Simply letting those two averages vary over the support of Y gives sharp bounds on the population average causal effect. Under our assumptions, these bounds are equal to both the Balke–Pearl and the Robins–Manski bounds.

The bioequivalence example discussed by Robins and Greenland is a complicated one. It is clear that with four qualitatively different treatments, randomization of a single binary assignment is generally not sufficient to identify average treatment effects for any of the four. A re-

lated but slightly simpler problem is that of partial compliance, where the binary assignment is to take a placebo or a full dose of the treatment, but individuals in the trial may take a partial dose (e.g., Efron and Feldman 1989). In related work (Angrist and Imbens 1995) we showed that in the RCM framework one can extend the assumptions made here to identify a weighted average of the slopes of the dose–response curves.

MOFFITT

Moffitt makes three main points and two minor points. He finds our choice of example unhelpful and our discussion of the literature on heterogeneous treatment effects lacking, and he offers an interpretation and some intuition for IV methods as a type of aggregation.

Moffitt also remarks that in the draft lottery example the assignment was not a true randomized clinical trial because the randomization was linked to birth dates, and incorrectly suggests this may affect the validity of the estimation of the complier average causal effect. Randomization does not have to be at the unit level. Given the stable unit treatment value assumption (SUTVA), the specific form of clustering present in this design does not affect the validity of the instrument in the presence of the seasonality effects Moffitt mentions, or of any other effects of birth dates on outcomes. We note, however, that in principle there are effects of the clustering on the precision of estimation.

In another remark, Moffitt claims that IV assumptions cannot be tested. Our independence assumptions (see also our discussion of Heckman's comment) do in fact impose restrictions on the joint distribution of the observables, as we discuss in Imbens and Rubin (1994). See also Balke and Pearl (1993) and Pearl (1996), who discuss the incompatibility of certain distributions with the IV assumptions given in this article.

Moffitt's first main point, regarding the choice of example, has been discussed in the Introduction to our rejoinder. Moffitt's second point concerns the treatment of heterogeneous effects in the econometric literature. A number of authors, including Heckman and Robb (1985) and Heckman (1990), have indeed discussed the estimation of models with heterogeneous effects using IV. However, the identification conditions they present generally are too strong to be useful in practice. A key condition of Heckman (1990) is the requirement that the support of the instrument cover the entire real line (whereas the instrument in our veterans application is a binary variable).

Moffitt also mentions Bjorklund and Moffitt (1987) as modeling heterogeneous treatment effects. This is an interesting application formulating the causal effects in terms of potential outcomes, although the specific model relies heavily on functional form and distributional assumptions rather than instruments to achieve identification.

Finally, Moffitt offers an additional example that provides an alternative interpretation for IV methods as aggregating within subpopulations defined by the instrument. This interpretation is useful, and in particular the analysis of variance (ANOVA) analogy to the difference between least squares regression and IV estimation offers an interesting perspective. However, Angrist (1991) has already discussed the grouping interpretation of IV estimators and given historical background for this idea, which dates back to Durbin (1954) and Friedman (1957). In addition, we do not find Moffitt's specific example of grouping very convincing. It is not clear why a comparison of earnings by city has a causal interpretation in this example. Because the candidate instrument in this case is closer to an attribute than a cause (in the Holland 1986 and Cox 1986 sense), the causal interpretation of the resulting IV estimate will be correspondingly weak.

ROSENBAUM

Rosenbaum makes some very interesting suggestions regarding alternative methods for inference. He also extends our sensitivity analysis to cover sensitivity to nonignorable assignment of the instrument. Both parts of his comment are welcome contributions to the discussion of IV methods. One issue that has limited the dialogue between economists and statisticians is the fact that econometric simultaneous equations models were not perceived as being interesting or relevant by the vast majority of statisticians. One of our goals here was to make at least some of these models accessible to the wider community of statisticians and to stimulate their contributions. We view Rosenbaum's comment as an early payoff to this effort.

Rosenbaum's first point concerns the lack of robustness of means. He suggests using more robust estimators such as the Hodges–Lehman estimator. It is interesting to note that despite the proliferation of discussions of median regression and more generally quantile regression as alternatives to mean regression in econometrics (e.g., Buchinsky 1994, Chamberlain 1994), there has been little work on robust alternatives for moment-based instrumental variables techniques (exceptions are Amemiya 1982 and Powell 1983). Clearly, the lack of robustness that motivated median regression as an alternative to mean regression applies equally well to IV problems. Rosenbaum's suggestion of the Hodges–Lehman estimator is novel and clearly deserves further attention.

Rosenbaum also points out that in typical economic applications the instrument is unlikely to be completely random, echoing Moffitt's point about our example of the draft lottery as an instrument not being representative of applications of IV methods in economics. This point is clearly valid and bolsters the case for a sensitivity analysis of the type Rosenbaum has suggested, here and in earlier work (e.g., Rosenbaum and Rubin 1983).

Finally, Rosenbaum points out that in his analysis of the Hodges–Lehman estimator, one need not make assumptions about the effect of the treatment for those who ignore encouragement. This important point can be made even stronger. One need not even define $Y_i(1)$ for never-takers, and similarly one need not define $Y_i(0)$ for always-takers. All we need to assume for units with $D_i(0) = D_i(1)$ is equality of the two potential outcomes; that is, $Y_i(0, D_i(0)) = Y_i(1, D_i(1))$, a weak form of the exclusion restriction.

CONCLUSION

The comments on our article cover a wide range of opinions, partly reflecting the gap between competing paradigms for evaluation research in statistics and econometrics. We believe that the gap between the two approaches can be narrowed. We hope that our article will make statisticians more appreciative of the insights offered by the IV framework invented by econometricians, while making economists more aware of the benefits of causal inference conducted in the potential outcomes framework developed by statisticians.

ADDITIONAL REFERENCES

Amemiya, T. (1982), "Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689–711.

Angrist, J. D. (1991), "Grouped-Data Estimation and Testing in Simple Labor-Supply Models," *Journal of Econometrics*, 47, 243–266.

Belson, W. A. (1956), "A Technique for Studying the Effects of a Television Broadcast," *Applied Statistics*, V, 195–202.

Bloom, H. (1984), "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 8, 225–246.

Buchinsky, M. (1994), "Changes in the U.S. Wage Structure 1963–1987: Application of Quantile Regression," *Econometrica*, 62, 405–458.

Chamberlain, G. (1994), "Quantile Regression, Censoring and the Structure of Wages," in *Advances in Econometrics*, ed. Simms, New York: Cambridge University Press.

Cochran, W. G. (1969), "The Use of Covariance in Observational Studies," *Applied Statistics*, 18, 270–275.

Cox, D. R. (1986), Comment on "Statistics and Causal Inference," by Holland, *Journal of the American Statistical Association*, 81, p. 963.

Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 3, 508–532.

Friedman, M. (1957), *A Theory of the Consumption Function*, Princeton, NJ: Princeton University Press.

Granger, C. (1986), Comment on "Statistics and Causal Inference," by Holland, *Journal of the American Statistical Association*, 81, p. 967–968.

Heckman, J. (1991), "Randomization and Social Policy Evaluation," Technical Working Paper 107, National Bureau of Economic Research.

Hendry, D., and Morgan, M. (1995), *The Foundations of Econometric Analysis*, Cambridge, U.K.: Cambridge University Press.

Pearl, J. (1996), "Causal Diagrams for Empirical Research," *Biometrika*, forthcoming.

Peters, C. C. (1941), "A Method of Matching Groups for Experiment With no Loss of Population," *Journal of Educational Research*, 34, 606–612.

Powell, J. (1983), "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 51, 1569–1575.

Rubin, D. B. (1973a), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.

——— (1973b), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 184–203.