



## Comment

Robert A. Moffitt

To cite this article: Robert A. Moffitt (1996) Comment, Journal of the American Statistical Association, 91:434, 462-465, DOI: [10.1080/01621459.1996.10476905](https://doi.org/10.1080/01621459.1996.10476905)

To link to this article: <https://doi.org/10.1080/01621459.1996.10476905>



Published online: 27 Feb 2012.



Submit your article to this journal [↗](#)



Article views: 48



View related articles [↗](#)



Citing articles: 1 View citing articles [↗](#)

- Models of Labor Force Dynamics," *Journal of Econometrics*, 18, 115–168.
- Gronau, R. (1974), "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.
- Hansen, L., and Sargent, T. (1991), *Rational Expectations Econometrics*, Boulder, CO: Westview Press.
- Heckman, J. (1974), "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 46, 695–712.
- (1978), "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 695–712.
- (1995), "Instrumental Variables: A Cautionary Tale," in *Essays in Honor of Sar Levitan*, eds. G. Mangum and S. Mangum, Kalamazoo, MI: W.E. Upjohn Press.
- Heckman, J., and Robb, R. (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in *Drawing Inferences From Self-Selected Samples*, ed. H. Wainer, Berlin: Springer-Verlag.
- Heckman, J., and Honoré, B. (1990), "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1150.
- (1989), "The Identifiability of the Competing Risks Model," *Biometrika*, 76, 325–330.
- Lucas, R., and Sargent, T. (1981), "Introduction," in *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press.
- McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, pp. 105–142.

## Comment

Robert A. MOFFITT

There are many unfortunate barriers to effective communication between statisticians and economists. The method of instrumental variables (IV) and associated methods for simultaneous equations and for "structural" estimation constitute one of the greatest. These methods are in the toolkit of virtually every economist and are among the most widely used techniques in the field. IV is discussed in every econometrics textbook, and three chapters of the 1984 *Handbook of Econometrics* are devoted to advanced IV and related issues, including nonlinear models (Amemiya 1984; Hausman 1984; Hsiao 1984). IV is widely regarded by economists as one of the most versatile and flexible of techniques, applicable in an enormous number of disparate applications. Yet it is scarcely used or discussed by statisticians, who often do not see the point of it all.

In this context, the attempt by Angrist, Imbens, and Rubin (AIR) to translate IV into terms that may be more understandable by statisticians must be welcomed. AIR translate IV into two frameworks familiar to statisticians. One is the well-known Rubin causal model (RCM). I find this translation to be correct and entirely appropriate, and hope that it is useful to statisticians. The other framework is the intention-to-treat (ITT) framework, with which statisticians are also quite familiar. I find this framework to have advantages as well as disadvantages. On the one hand, the noncompliance problem that is at the heart of the ITT framework is a nice illustration of the econometric problem of "endogeneity" that leads to IV estimation in economics. The notion that the difference in means between experimentals and controls should be inflated by the difference in the percentage treated in the two groups is also common to the ITT and IV frameworks. On the other hand, ITT analysis is conventionally discussed in the context of a randomized clinical trial (RCT), and AIR do so as well. This provides by necessity

an obvious and convincing instrument—the experimental treatment assignment. (For another discussion of experimental assignment as an IV, see Heckman, in press.) Yet in the vast majority of work in economics, observational data are used instead, and consequently, some of the assumptions of IV stated by AIR—the random assignment assumption and the exclusion restriction—play a far more critical role in applied work than is suggested by the ITT–RCT framework.

In what follows, I make a few remarks about IV from the viewpoint of an economist, in hopes of further illuminating the interpretation and breadth of the technique. (For other recent work by economists on IV and identification in related models, see Bound, Jaeger, and Baker 1995; Manski 1990, 1994; and Staiger and Stock 1993.)

### Heterogeneous Response Interpretation

The simultaneous equations model given by AIR in their equations (1)–(3) is one of the models considered in the first attempt at a comprehensive econometric treatment of the causal effects problem by Heckman and Robb (1985, 1986). That work was in turn based on the original formulation of the dummy endogenous variable model by Heckman (1978) (on which the Maddala, Bowden–Turkington, and Heckman–Robb papers cited by AIR in Section 2 are based). Although AIR find the use of unobservables in the specification of equations (1)–(3) and the assumptions surrounding it to be nonintuitive, it is important to stress that the model in those equations is nevertheless directly translatable into, and is equivalent to, the Rubin causal model (RCM) with one modification: to allow the treatment effect in equation (1) to vary across individuals; for example,  $\beta_1(i)$ . With that modification,  $\beta_1(i) = Y_i(1) - Y_i(0)$  is the Rubin causal effect of  $D_i$  on  $Y_i$  given by AIR in Definition 2. The assumptions of linearity, additive disturbances, and other aspects of the specification in equations (1)–(3) are entirely unrestrictive in this simple a model. Thus, as

Robert A. Moffitt is Professor of Economics, Department of Economics, Johns Hopkins University, Baltimore, MD 21218. An earlier version of these remarks was presented at the meetings of the American Statistical Association, Orlando, Florida, August 13–17, 1995. The author would like to thank James Heckman, V. Joseph Hotz, and Charles Manski for comments.

AIR note, the issue is which framework provides the better intuition. Of course, one should not expect economists and statisticians, or even different individuals within each discipline, to find their intuition in the same way, and there is no reason not to have the model translated into multiple frameworks.

While the constant effect assumption is made in most IV work in economics as a whole, the heterogeneous effect model nevertheless has a long history in certain areas. For example, heterogeneous effects appear in the basic multinomial discrete choice model in relative preferences for different alternatives (McFadden 1974, 1984). The switching regression model of Heckman (1978) and Lee (1979), which is closely related to the comparative advantage model of Roy (1951), has heterogeneous response to “regime switching” as a key characteristic. In the treatment effects literature, Heckman and Robb (1985, 1986) made the heterogeneous effect model explicit in their analysis and discuss its IV estimation. Björklund and Moffitt (1987) conducted an actual empirical analysis of a heterogeneous-effect model but estimated it with maximum likelihood instead of IV. There has been a steady stream of studies analyzing the heterogeneous response model since those papers.

To many economists, the heterogeneous coefficient formulation has great intuitive appeal because it assigns an explicit parameter to the main object of interest—the true treatment effect for each individual  $i$ . It also gives explicit representation to the potential, but unobserved, treatment effect of noncompliers, which AIR correctly emphasize is so important to the conceptualization of the problem. (They are a little misleading in suggesting that the concept of potential outcomes is missing in the econometric formulation of the problem; in fact, that concept is key to the econometric formulation as well, even if less explicitly stated.) In addition, such a parameterization gives the monotonicity assumption a ready illustration, for one possible embodiment of that assumption is a model in which treatment receipt for individual  $i$  is a monotonic function of  $\beta_1(i)$ ; for example, those in the experimental group who receive the treatment are those who are more affected by it.

#### Alternative Interpretations of IV

Although AIR find the specification of (1)–(3) and the statement of IV assumptions in terms of the unobservables in those equations to be nonintuitive, economists usually gain their intuition for IV from what they see to be the implications of the assumptions for specific applications. It is for this reason that the RCT framework does not well illustrate the source of intuition for economists, who generally use observational data.

In addition, the military lottery application discussed by AIR is not typical of most IV applications in economics, for most do not involve any explicit randomization. The lottery example is not a pure RCT in any case, because the randomization was based on an intervening variable—an individual characteristic (birthdate). Pure RCT’s instead randomize individuals directly into experimental and control groups. Consequently, the lottery application requires one additional assumption—birthdate does not directly af-

fect mortality—to satisfy the exclusion restriction and make IV possible. The fact that there are well-known seasonal effects in birth rates (Lam and Miron 1991) that may have a various health-related and socioeconomic antecedents and consequences suggests that the validity of this additional assumption cannot be immediately accepted without further investigation.

A more typical, perhaps even prosaic, economic example, and one that provides an alternative interpretation and source of intuition for IV, is the following. Suppose that we wish to estimate the effect of a job training program on future earnings, and we have data from two different cities on the earnings of men who have and have not gone through job training at some point in the past. Comparing the earnings of trained and untrained workers within each city alone, or pooling the data from both cities and making the same comparison, would yield poor estimates of the effect of training if the untrained workers were different from the trained workers even if they had not gone through training; that is, if there is nonignorable selection bias. But if, say, city A had more funds and offered more training slots than city B, then the fraction of workers who have been trained will be higher in city A than in city B. Consequently, the effect of training could be estimated by regressing mean earnings in each city—the mean taken over trained and untrained workers combined—on the fraction of workers in the city who were trained. The resulting regression coefficient is simply the IV estimate given by AIR in their equation (6):

$$\beta_1^{IV} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}, \quad (1)$$

where  $\bar{Y}_Z$  and  $\bar{D}_Z$  are the means of  $Y$  and  $D$  for the groups  $Z = 1$  and  $Z = 0$ . Thus the difference in means between the two populations is inflated by the change in the fraction “treated,” exactly as in the ITT framework.

The econometric assumptions necessary for this estimator to represent a causal effect in the context of AIR’s equations (1)–(3) have a ready intuition in the context of this example. The random assignment assumption  $E(Z_i \varepsilon_i) = 0$  (i.e., mean independence) is just the assumption that the greater number of training slots in city A is “exogenous”; that is, unrelated to earnings in the absence of training. Put differently, it is the assumption that the across- $Z$  variation represents a true “experiment.” This assumption would be violated, for example, if city A obtained more training funds because its workers had lower earnings than those in city B in the first place. The exclusion restriction—which is implicit in equations (1)–(3) because  $Z_i$  does not appear in equation (1)—is just the assumption that there is no direct effect of city of residence on earnings. This assumption would be violated if, for example, the labor market in one city was healthier than that in the other city, which would make earnings different even in the absence of training differences. The monotonicity assumption is just the assumption that everyone who received training in city B would receive training if they resided in city A.

This example provides an alternative interpretation of IV, as simply representing a *comparison in a different*

*dimension*—in this case, a comparison across cities (i.e., across values of  $Z_i$ ) instead of a comparison of trained and untrained workers within cities (i.e., across individual values of  $D_i$ ). Expressing the two methods as simply comparisons along different dimensions puts them on a more equal footing and leads to the additional observation that either could be correct or incorrect (or neither could be, of course). The across-city IV comparison would be biased if the allocation of city funds were based on earnings, for example, as noted previously; but, providing that there is no selection of workers into training *within* each city, a least squares regression that includes a city dummy (i.e., conditions on  $Z_i$ ) would yield unbiased treatment effect estimates. In the econometric literature, Goldberger (1972) was the first to note this point, showing that if selection into treatment status is based only on an auxiliary variable  $Z_i$ , then one need only condition on that variable to obtain consistent and unbiased treatment effects (even if the coefficient on that auxiliary variable itself is biased). The later econometric literature clarified the distinction between this case—selection on observables—and the case of selection on unobservables.

Much of the debate in economics involves arguments in specific empirical applications over whether a particular instrument  $Z_i$  does or does not improve the estimate of treatment effects, given that it can make things worse as well as better. Making such a determination is particularly difficult when it is recognized that no statistical test or specification test can distinguish such models at this simple level. If the two estimates are different, whether one estimate is significantly different from the other can be tested only under the null that one is correct. Although AIR state that IV can be subjected to “sensitivity” testing, the fundamental IV assumptions cannot be tested if they are “just” identifying—that is, if they are a minimal and thus necessary rather than sufficient set of assumptions to obtain treatment effects. (See Heckman and Robb 1986, Heckman and Hotz 1989, and Moffitt 1989 for discussions of this important point.)

The city example provides yet another interpretation of IV, which is as a *method of aggregation* (Moffitt, 1996). The IV estimator represents a least squares regression using aggregates taken over  $Y_i$  and  $D_i$  within cells of  $Z_i$ . A related intuition is based on an analysis of variance (ANOVA) analogy, for the IV estimator uses the covariance of  $Y_i$  and  $D_i$  “between” cities rather than “within” cities. Indeed, it is easy to show that the ordinary least squares (OLS) estimate of  $\beta_1$  (i.e., the estimate obtained by comparing treatments and comparisons in the total, pooled sample) is a weighted average of the IV (between) estimator and the within estimator:

$$\beta_1^{\text{OLS}} = k\beta_1^{\text{IV}} + (1 - k)\beta_1^{\text{W}} \quad (2)$$

where  $k$  is the fraction of the total variance of  $D$  that arises from the “between” and  $\beta_1^{\text{W}}$  is the treatment effect based on the within variation (i.e., the coefficient on  $D_i$  in a regression of  $Y_i$  on  $D_i$  and a  $Z_i$  dummy). The exact decomposition

shown in (2) assumes that the sample size is the same in all cities.

The ANOVA analogy can also be used to relate IV to the propensity score method of Rosenbaum and Rubin (1983). In the simple case of a single dummy variable  $Z_i$ , conditioning on the propensity score is identical to conditioning on  $Z_i$  and hence is equivalent to the within estimator,  $\beta_1^{\text{W}}$ . The IV estimator, on the other hand, can be shown to be equivalent to that obtainable by regressing  $Y_i$  on the propensity score itself; that is, by *replacing* the treatment dummy  $D_i$  by the propensity score. (This is the two-stage least squares version of IV.) This yields a treatment effect estimate based on the “between.”

## ADDITIONAL REFERENCES

- Amemiya, T. (1984), “Nonlinear Regression models,” in *Handbook of Econometrics*, Vol. I, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 333–389.
- Björklund, A., and Moffitt, R. (1987), “Estimation of Wage Gains and Welfare Gains in Self-Selection Models,” *Review of Economics and Statistics*, 69, 42–49.
- Bound, J., Jaeger, D., and Baker, R. (1995), “Problems with Instrumental Variables When the Correlation Between the Instruments and the Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90, 443–450.
- Goldberger, A. (1972), “Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations,” Discussion Paper 123-72, University of Wisconsin, Institute for Research on Poverty.
- Hausman, J. (1984), “Specification and Estimation of Simultaneous Equations Models,” in *Handbook of Econometrics*, Vol. I, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 392–448.
- Heckman, J. (1978), “Dummy Endogenous Variables in a Simultaneous Equation System,” *Econometrica*, 46, 931–960.
- (in press), “Randomization as an Instrumental Variable,” *Review of Economics and Statistics*.
- Heckman, J., and Hotz, V. J. (1989), “Choosing Among Alternative Non-experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training,” *Journal of the American Statistical Association*, 84, 862–874.
- Heckman, J., and Robb, R. (1985), “Alternative Models for Evaluating the Impact of Interventions,” in *Longitudinal Analysis of Labor Market Data*, eds. J. Heckman and B. Singer, Cambridge, U.K.: Cambridge University Press, pp. 156–245.
- (1986), “Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes,” in *Drawing Inferences from Self-Selected Samples*, ed. H. Wainer, New York: Springer-Verlag, pp. 63–107.
- Hsiao, C. (1984), “Identification,” in *Handbook of Econometrics*, Vol. I, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 224–283.
- Lam, D., and Miron, J. (1991), “Seasonality of Births in Human Populations,” *Social Biology*, 38, 51–78.
- Lee, L. F. (1979), “Identification and Estimation in Binary Choice Models With Limited (Censored) Dependent Variables,” *Econometrica*, 47, 977–996.
- Manski, C. (1990), “Nonparametric Bounds for Treatment Effects,” *American Economic Review*, 80, 319–323.
- (1994), “The Selection Problem,” in *Advances in Econometrics, Sixth World Congress*, ed. C. Sims, Cambridge, U.K.: Cambridge University Press, pp. 143–170.
- McFadden, D. (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, pp. 105–142.
- (1984), “Econometric Analysis of Qualitative Response Models,” in *Handbook of Econometrics*, eds. Z. Griliches and M. Intriligator, Amsterdam: North-Holland, pp. 1396–1457.
- Moffitt, R. (1989), Comment on “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The

Case of Manpower Training," by J. Heckman and V. J. Hotz, *Journal of the American Statistical Association*, 84, 877-878.

——— (1996), "Selection Bias Adjustment in Treatment-Effect Models as a Method of Aggregation," in *1995 Proceedings of the American Statistical Association*.

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity

Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Roy, A. D. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135-146.

Staiger, D., and Stock, J. (1993), "Instrumental Variable Regression With Weak Instruments," mimeo, Harvard University.

## Comment

Paul R. ROSENBAUM

### 1. INTRODUCTION

Angrist, Imbens, and Rubin (AIR) deserve congratulations for a wonderful paper. Linking econometrics with experimental design, they have illuminated both fields. I particularly admire the care they take in defining estimands with few modeling assumptions, in stating assumptions in tangible terms, and in examining the appropriateness and consequences of those assumptions.

In this comment, I would like to slightly restate their argument in terms of an artificial example, then generalize the argument to a larger class of estimators (the Hodges-Lehmann estimators), briefly indicate how one can conduct a sensitivity analysis in a nonrandomized study, and conclude with an observation about the case in which some subjects have unalterable treatment assignments.

### 2. AN ARTIFICIAL EXAMPLE: ENCOURAGING EXERCISE FOR LUNG DISEASE

The following artificial example illustrates and restates several of the points made by AIR, but its main purpose is to aid in Section 3 in discussing a generalization of the instrumental variables (IV) estimate. Table 1 describes a randomized experiment with 10 subjects suffering from chronic obstructive lung disease (COLD), of whom 5 were randomly selected and encouraged to exercise. So the randomization determines who was encouraged to exercise; that is,  $Z_i$ . In fact, not all subjects complied, as indicated by  $D_i(Z_i)$ . Subjects 1, 2, and 3 exercised as they were encouraged to do, but subjects 4 and 5 ignored the encouragement and did not exercise. Subject 6 was not encouraged to exercise ( $Z_6 = 0$ ) but did so anyway ( $D_6(0) = 1$ ). The outcome is forced expiratory volume (FEV), a measure of lung function, larger values indicating better health, recorded on a convenient integer scale. The quantity  $Y_i(0)$  is the outcome that would have been observed from subject  $i$  in the absence of exercise. As in AIR's exclusion restriction and in Holland's (1988) discussion of encouragement designs, it is exercise that may have an effect, but encouragement has an effect only if it influences exercise. In Table 1 exercise raises the outcome, FEV, by 3 units for all subjects; that is, the ob-

served response from subject  $i$  is  $\tilde{Y}_i = Y_i(0) + 3D_i(Z_i)$  (e.g.,  $\tilde{Y}_1 = Y_1(0) + 3D_1(1) = 5 + 3 = 8$ ).

Several features of Table 1 are of note. First, the FEV response that would be observed from subject  $i$  in the absence of exercise,  $Y_i(0)$ , is unaffected by encouragement  $Z_i$  or exercise  $D_i(Z_i)$ , and in the theory of randomized experiments,  $Y_i(0)$  is a fixed feature of subject  $i$  not varying with the random assignment of encouragement  $Z_i$ . By good fortune in this artificial example, the distribution of  $Y_i(0)$  is perfectly balanced in encouraged ( $Z_i = 1$ ) and control ( $Z_i = 0$ ) groups. Randomization produces such balance in expectation in randomized experiments of all sizes, and in large experiments approximate balance is likely, but the exact balance in Table 1 is an unnecessary but tidy convenience useful in exposition. In short, the randomization worked—without treatment, the two randomized groups ( $Z_i = 1$ ) and ( $Z_i = 0$ ) would have had similar outcomes. The observed responses  $\tilde{Y}_i$  are not balanced of course, because encouragement  $Z_i$  increases exercise  $D_i(Z_i)$ , which increases  $\tilde{Y}_i$ . Notice also that healthy subjects are more likely to exercise. More precisely, subjects who would have had high FEV absent exercise—subjects with high  $Y_i(0)$ —are more likely to have  $D_i(Z_i) = 1$ . Encouragement appears to increase the amount of exercise, but subjects with low  $Y_i(0)$  do not exercise even if encouraged.

The traditional advice in randomized clinical trials is that the groups formed by randomization should be compared; here, the encouraged ( $Z_i = 1$ ) and control ( $Z_i = 0$ ) groups. The difference in means is  $(8 + 7 + 6 + 2 + 1)/5 - (8 + 4 + 3 + 2 + 1)/5 = 6/5 = 1.2$ . This is a sensible estimate of the effect of encouragement. Exercise raises FEV by 3, but encouragement raises it only by 1.2 on average, because many subjects do not exercise when encouraged and some exercise without encouragement. A mistaken estimate of the effect of exercise compares those who exercised ( $D_i(Z_i) = 1$ ) to those who did not ( $D_i(Z_i) = 0$ )—namely,  $(8 + 7 + 6 + 8)/4 - (2 + 1 + 4 + 3 + 2 + 1)/6 = 7.250 - 2.167 = 5.083$ . This estimate grossly overstates the effect of exercise because healthier subjects were more likely to exercise. The instrumental variables estimate starts with