Identification of Causal Effects Using Instrumental Variables

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

We outline a framework for causal inference in settings where assignment to a binary treatment is ignorable, but compliance with the assignment is not perfect so that the receipt of treatment is nonignorable. To address the problems associated with comparing subjects by the ignorable assignment—an "intention-to-treat analysis"—we make use of instrumental variables, which have long been used by economists in the context of regression models with constant treatment effects. We show that the instrumental variables (IV) estimand can be embedded within the Rubin Causal Model (RCM) and that under some simple and easily interpretable assumptions, the IV estimand is the average causal effect for a subgroup of units, the compliers. Without these assumptions, the IV estimand is simply the ratio of intention-to-treat causal estimands with no interpretation as an average causal effect. The advantages of embedding the IV approach in the RCM are that it clarifies the nature of critical assumptions needed for a causal interpretation, and moreover allows us to consider sensitivity of the results to deviations from key assumptions in a straightforward manner. We apply our analysis to estimate the effect of veteran status in the Vietnam era on mortality, using the lottery number that assigned priority for the draft as an instrument, and we use our results to investigate the sensitivity of the conclusions to critical assumptions.

KEY WORDS: Compliers; Intention-to-treat analysis; Local average treatment effect; Noncompliance; Nonignorable treatment assignment; Rubin-Causal-Model; Structural equation models.

1. INTRODUCTION

Economists are typically interested in estimating causal effects rather than mere associations between variables. Potentially interesting causal effects include the effects of education on employment and earnings, the effects of employment training programs on subsequent labor market histories, and the effects of a firm's inputs on its output. The dominant approach to making inferences about causal effects in economics over the last four decades is based on *structural equation models,* which rely on the specification of systems of equations with parameters and variables that attempt to capture behavioral relationships and specify the causal links between variables. Goldberger (1972) and Morgan (1990) provided historical perspectives on these models, which date back to Wright (1928, 1934) and Haavelmo (1943, 1944). Inference in structural equation models often exploits the presence of *instrumental variables* (IV). These are variables that are explicitly excluded from some equations and included in others, and therefore correlated with some outcomes only through their effect on other variables.

Rather than relying on structural equation models, causal inference in statistics, going back at least to work by Fisher (1918, 1925) and Neyman (1923) on agricultural experiments, is fundamentally based on the randomized experiment (see also Kempthorne 1952 and Cox 1958). The basic notion in this formulation, which has been extended by Rubin (1974, 1978) to more complicated situations, including observational studies without randomization, is that of *potential outcomes.* The causal effect of a treatment on a single individual or unit of observation is the comparison (e.g., difference) between the value of the outcome if the unit is treated and the value of the outcome if the unit is not treated. The target of estimation, the estimand, is typically the average causal effect, defined as the average difference between treated and untreated outcomes across all units in a population or in some subpopulation (e.g., males or females). For this definition of causality to be applicable to samples with units already exposed to treatments, we must be able to imagine observing outcomes on a unit in circumstances other than those to which the unit was actually exposed. This approach is now widely used in statistics and epidemiology (e.g., Efron and Feldman 1991 and Greenland and Robins 1986), where it is often referred to as the Rubin Causal Model (RCM; Holland [1986]).

In this article we provide a link between these approaches, capitalizing on the strengths of each. Earlier work combining elements of these approaches includes studies by Hearst, Newman, and Hulley (1986), Holland (1988), Permutt and Hebel (1989), Sommer and Zeger (1991), and Imbens and Angrist (1994). We show how the IV estimand can be given a precise and straightforward causal interpretation in the potential outcomes framework, despite nonignorability of treatment received. This interpretation avoids drawbacks of the standard structural equation framework, such as constant effects for all units, and delineates critical assumptions needed for a causal interpretation. The IV approach provides an alternative to a more conventional intention-to-treat analysis, which focuses solely on the average causal effect of assignment on the outcome (Lee, Ellenberg, Hirtz, and Nelson 1991).

As we show in the context of a specific application, our formulation of these assumptions makes it easier for researchers to judge whether or not a causal interpretation of the instrumental variables estimand is plausible. Standard

Joshua D. Angrist is Associate Professor, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. Guido W. Imbens is Associate Professor, Department of Economics, and Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. The authors thank Gary Chamberlain, Clifford Clogg, Sander Greenland, Judy Hellerstein, Paul Rosenbaum, the editor, an associate editor and several referees for comments, the National Science Foundation for financial support, and Tom Newman and Norman Hearst for sharing their data.

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

IV procedures rely on judgments regarding the correlation between functional-form-specific disturbances and instruments. In contrast, our approach forces the researcher to consider the effect of exposing units to specific treatments. If it is not possible (or not plausible) to envision the alternative treatments underlying these assumptions, the use of these techniques may well be inappropriate. Moreover, by separating and defining the critical assumptions, our formulation allows for a clear assessment of the consequences of violations of these assumptions through sensitivity analysis under more general models. Our main results are summarized in three propositions: the first provides conditions for a causal interpretation of the IV estimand, and the others reveal the consequences of violations of the critical assumptions.

We develop our presentation in the context of an evaluation of the effect of serving in the military on health outcomes. Data for this study come from the Vietnam era, when priority for conscription was randomly allocated through the draft lottery. For expository purposes, and to be precise without cumbersome notation, we use the simplest possible example: both the "treatment" (i.e., serving in the military or not, denoted by D) and the "assignment" (i.e., draft status, determined by lottery number, denoted by Z) are binary. If compliance with the draft had been perfect, then all those with a low lottery number ($Z = 1$) would have served in the military ($D = 1$), and all those with a high lottery number ($Z = 0$) would not have served ($D = 0$). We assume that we observe values of Z, D, and the health outcome Y for each person. Our basic results, however, are not limited to this case with binary treatment and binary instrument. The approach developed here can be extended to multi-valued treatments and instruments as in Angrist and Imbens (1995) and Angrist, Graddy and Imbens (1995). Moreover, the generalization to cases with covariates is, in principle, immediate by applying our results at distinct values of the covariates. Also, fully principled methods of estimation using likelihood-based or Bayesian techniques can be derived as in Imbens and Rubin (1994a).

In Section 2 we briefly describe the structural equation approach to causal inference in economics. In Section 3 we develop an alternative approach based on the RCM, and the approaches are contrasted in Section 4. In Section 5 we discuss how to evaluate the sensitivity of the IV estimand to two of the critical assumptions presented in Section 3. In Section 6 we apply this approach to our draft lottery example, where we formulate the critical assumptions in the RCM framework and investigate the implications of violations of these assumptions.

2. STRUCTURAL EQUATION MODELS IN ECONOMICS

Following Goldberger (1972), we define structural equation models as "stochastic models in which each equation represents a causal link, rather than a mere empirical association" (p. 979). Such models are widely used in economics, going back to work by Wright (1928, 1934), Schultz (1928), and Haavelmo (1943, 1944).

A structural equation model for the problem of inferring the effect of veteran status on a health outcome is the *dummy endogenous variable model* (see, e.g., Maddala 1983; Bowden and Turkington 1984; Heckman and Robb 1985). For person i, let Y_i be the observed health outcome, let D_i be the observed treatment (i.e., veteran status), and let Z_i be the observed draft status. A standard dummy endogenous variables model for this problem would have the form

$$Y_i = \beta_0 + \beta_1 \cdot D_i + \varepsilon_i, \tag{1}$$

$$D_i^* = \alpha_0 + \alpha_1 \cdot Z_i + \nu_i \tag{2}$$

and

$$D_i = \begin{cases} 1 & \text{if } D_i^* > 0, \\ 0 & \text{if } D_i^* \leq 0. \end{cases} \tag{3}$$

In this model β_1 represents the causal effect of D on Y. Although simple, this model is typical of the econometric approach to discrete choice (in this case, the choice to serve in the military or not). The latent index formulation involving D_i^* originates in the notion that compliance is a choice determined by comparison of the expected utility of serving and not serving. We note that this dummy endogenous variables model shares many features with the classical simultaneous equations model (Haavelmo 1943): an underlying linear structure, constant coefficients, and a reliance on error terms to characterize omitted variables.

The first assumption typically invoked to identify β_1 is that Z_i is uncorrelated with the disturbances ε_i and ν_i:

$$E[Z_i \cdot \varepsilon_i] = 0, \qquad E[Z_i \cdot \nu_i] = 0. \tag{4}$$

The assumption that the correlation between ε and Z_i is zero and the absence of Z in Equation (1) captures the notion that any effect of Z on Y must be through an effect of Z on D. This is a key assumption in econometric applications of instrumental variables. A second assumption is that the covariance between the treatment D_i and assignment Z_i differs from zero; that is,

$$\operatorname{cov}(D_i, Z_i) \neq 0, \tag{5}$$

which can be interpreted as requiring that α_1 differ from zero. If Z_i satisfies these two assumptions, then it is considered an IV in this model. In general D_i, the *endogenous regressor* in econometric terminology, is potentially correlated with ε_i because the two disturbances ε_i and ν_i are potentially correlated. This implies that the receipt of treatment D_i is not *ignorable* (Rubin 1978) and, in econometric terminology, not *exogenous.*

For this simple example, the IV estimator is defined as the ratio of sample covariances (Durbin 1954)

$$\begin{aligned}\hat{\beta}_1^{\mathrm{IV}} &= \widehat{\operatorname{cov}}(Y_i, Z_i)/\widehat{\operatorname{cov}}(D_i, Z_i) \\ &= \frac{\sum_{i=1}^N Y_i Z_i/\sum_{i=1}^N Z_i - \sum_{i=1}^N Y_i(1-Z_i)/\sum_{i=1}^N (1-Z_i)}{\sum_{i=1}^N D_i Z_i/\sum_{i=1}^N Z_i - \sum_{i=1}^N D_i(1-Z_i)/\sum_{i=1}^N (1-Z_i)},\end{aligned} \tag{6}$$

where the last equality follows from the binary nature of the instrument.

Structural equation models such as Equations (1)–(3) have not found widespread use among statisticians. One reason is the sensitivity of these models to critical assumptions (see Little 1985) and their apparent inability to reproduce experimental results (see Lalonde 1986). Another reason is the fact that critical assumptions are cast in terms of disturbances from incompletely specified regression functions (i.e., ε_i and ν_i), rather than in terms of intrinsically meaningful and potentially observable variables. Typically the researcher does not have a firm idea what these disturbances really represent, and therefore it is difficult to draw realistic conclusions or communicate results based on their properties. The focus of this article is on the causal interpretation of the limit of the estimator in Equation (6); that is, the *IV estimand,* using the potential outcomes framework, and on the formulation of the critical assumptions in a more transparent manner to make these models more accessible to statisticians.

3. CAUSAL ESTIMANDS WITH INSTRUMENTAL VARIABLES

In this section, we set out an alternative framework for a causal interpretation of the IV estimand based on potential outcomes. First, we discuss the RCM approach to analyzing the causal effects of assignment on treatment received and on the outcome of interest (the intention-to-treat effects). We then define the causal effect of interest, that of treatment received on the outcome, in terms of potential outcomes. Finally, we show how the IV estimand links the two average intention-to-treat effects to a subpopulation average of the causal effect of interest.

3.1 The Rubin Causal Model

As before, $Z_i = 1$ implies that person i has a low lottery number (i.e., would potentially get called to serve in the military), whereas $Z_i = 0$ indicates that person i has a high lottery number (i.e., would not get called to serve in the military). The subsequent notation for D and Y is somewhat different from that in Section 2 because of the need to represent potential outcomes. Let $\mathbf{Z}$ be the N-dimensional vector of assignments with ith element Z_i, and let $D_i(\mathbf{Z})$ be an indicator for whether person i would serve given the randomly allocated vector of draft assignments $\mathbf{Z}$. In a world of perfect compliance with the draft, $D_i(\mathbf{Z})$ would equal Z_i for all i; that is, those with low lottery numbers would actually serve and none of those with high lottery numbers would serve. In practice, $D_i(\mathbf{Z})$ can differ from Z_i for various reasons: individuals may volunteer for military service, they may avoid the draft, or they may be deferred for medical or family reasons.

Similar to the definition of $D_i(\mathbf{Z})$, we define $Y_i(\mathbf{Z}, \mathbf{D})$ to be the response for person i given the vector of service indicators $\mathbf{D}$ and the vector of draft priorities $\mathbf{Z}; \mathbf{Y}(\mathbf{Z}, \mathbf{D})$ is the N vector with ith element $Y_i(\mathbf{Z}, \mathbf{D})$. We refer to $D_i(\mathbf{Z})$ and $Y_i(\mathbf{Z}, \mathbf{D})$ as "potential outcomes." The concept of potential outcomes used here can be viewed as analogous to Neyman's (1923) notion of "potential yields" in randomized agricultural experiments, as extended by Rubin (1974, 1978, 1990, 1991) to observational studies where the potential outcomes are partially revealed by a general treatment assignment mechanism, to situations with possible variation of treatments and with possible interference between units, and to Bayesian and likelihood inference where the potential outcomes and assignment have a joint probability distribution. As originally formulated, the potential outcomes $D_i(\mathbf{Z})$ and $Y_i(\mathbf{Z}, \mathbf{D})$ are fixed but unknown values partially observed through the assignment of treatments to units. Differences in these potential outcomes due to assigned and received treatments will be revealed by analyzing data obtained by randomly assigning $\mathbf{Z}$ in the finite population of N units under study. Our initial goal is to provide inferences solely about this finite population.

In evaluation research, some assumptions about how units interact and the variety of possible treatments are required. Our notation has already restricted both Z and D to have only two levels; that is, there is no partial compliance. Here we follow the convention in statistics and medical research by assuming no interference between units.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978, 1980, 1990).

a. If $Z_i = Z_i'$, then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$.

b. If $Z_i = Z_i'$ and $D_i = D_i'$, then $Y_i(\mathbf{Z}, \mathbf{D}) = Y_i(\mathbf{Z}', \mathbf{D}')$.

SUTVA implies that potential outcomes for each person i are unrelated to the treatment status of other individuals. This assumption allows us to write $Y_i(\mathbf{Z}, \mathbf{D})$ and $D_i(\mathbf{Z})$ as $Y_i(Z_i, D_i)$ and $D_i(Z_i)$ respectively. SUTVA is an important limitation, and situations where this assumption is not plausible cannot be analyzed using the simple techniques outlined here, although generalizations of these techniques can be formulated with SUTVA replaced by other assumptions.

Given the set of potential outcomes, we can define the causal effects of Z on D and on Y in the standard fashion (Rubin, 1974).

Definition 1: Causal Effects of Z on D and Z on Y.
The causal effect for individual i of Z on D is $D_i(1) - D_i(0)$. The causal effect of Z on Y is $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$.

In the context of a clinical trial with imperfect compliance these are the intention-to-treat effects, and we adopt this jargon here.

Although Bayesian or likelihood-based inference is straightforward if treatment assignment is ignorable, even if not completely random (Rubin 1978), we assume random assignment here to avoid tangential issues.

Assumption 2: Random Assignment.
The treatment assignment Z_i is random:

$$\Pr(\mathbf{Z} = \mathbf{c}) = \Pr(\mathbf{Z} = \mathbf{c}')$$

for all $\mathbf{c}$ and $\mathbf{c}'$ such that $\iota^T\mathbf{c} = \iota^T\mathbf{c}'$, where ι is the N-dimensional column vector with all elements equal to one.

Given SUTVA and random assignment, unbiased estimators for the average intention-to-treat effects can be obtained by taking the difference of sample averages of Y and D classified by the value of Z; that is, by treatment-control mean differences. This has been well known since at least Neyman (1923). Formally, the unbiased estimator for the average causal effect of Z on Y can be written as

$$\frac{\sum_i Y_i Z_i}{\sum_i Z_i} - \frac{\sum_i Y_i(1-Z_i)}{\sum_i (1-Z_i)}$$
$$= \frac{(1/N)\sum_{i=1}^N Y_i Z_i - (1/N)\sum_{i=1}^N Y_i \cdot (1/N)\sum_{i=1}^N Z_i}{(1/N)\sum_{i=1}^N Z_i Z_i - (1/N)\sum_{i=1}^N Z_i \cdot (1/N)\sum_{i=1}^N Z_i}, \tag{7}$$

and for the average causal effect of Z on D the unbiased estimator can be written as

$$\frac{\sum_i D_i Z_i}{\sum_i Z_i} - \frac{\sum_i D_i(1-Z_i)}{\sum_i (1-Z_i)}$$
$$= \frac{(1/N)\sum_{i=1}^N D_i Z_i - (1/N)\sum_{i=1}^N Y_i \cdot (1/N)\sum_{i=1}^N Z_i}{(1/N)\sum_{i=1}^N Z_i Z_i - (1/N)\sum_{i=1}^N Z_i \cdot (1/N)\sum_{i=1}^N Z_i}. \tag{8}$$

The ratio of (7) and (8) equals the conventional instrumental variables estimator (6). The limit of the IV estimator (i.e., the IV estimand), therefore equals the ratio of average intention-to-treat effects.

3.2 Instrumental Variables

The critical feature of the problem of evaluating a treatment under imperfect compliance is that even if assignment Z_i is random or ignorable, the actual receipt of treatment D_i is typically nonignorable. Therefore the difference of outcome averages by treatment received does not provide an unbiased or even consistent estimate of the average causal effect of D on Y. In fact, we require additional assumptions just to define the causal effect of D on Y in a meaningful way. The following assumption requires the treatment assignment to be unrelated to potential outcomes once treatment received is taken into account.

Assumption 3: Exclusion Restriction.
$\mathbf{Y}(\mathbf{Z},\mathbf{D}) = \mathbf{Y}(\mathbf{Z}',\mathbf{D})$ for all $\mathbf{Z},\mathbf{Z}'$ and for all $\mathbf{D}$.

This assumption implies that $Y_i(1,d) = Y_i(0,d)$ for $d = 0, 1$. It captures the notion underlying instrumental variables procedures that any effect of Z on Y must be via an effect of Z on D. Because the exclusion restriction relates quantities that can never be jointly observed, (i.e., $Y_i(0,d)$ and $Y_i(1,d)$), it is not directly verifiable from the data at hand although it has testable implications when combined with Assumptions 1 and 2. Imbens and Rubin (1994b) discussed a weaker version of the exclusion restriction that impose restrictions only on outcomes that can potentially be observed (i.e., $Y_i(z, D_i(z))$.)

By virtue of Assumption 3, we can now define potential outcomes $\mathbf{Y}(\mathbf{Z},\mathbf{D})$ as a function of $\mathbf{D}$ alone:

$$\mathbf{Y}(\mathbf{D}) = \mathbf{Y}(\mathbf{Z},\mathbf{D}) = \mathbf{Y}(\mathbf{Z}',\mathbf{D}) \quad \forall\, \mathbf{Z},\mathbf{Z}' \quad \text{and} \quad \forall\, \mathbf{D},$$

and then by Assumption 1 we can write $Y_i(D_i)$ instead of $Y_i(\mathbf{Z},\mathbf{D})$.

We now have notation for the causal effects of interest.

Definition 2: Causal Effects of D on Y.
The causal effect of D on Y for person i is $Y_i(1) - Y_i(0)$.

Although we can never observe any of these causal effects, for people with $D_i(0) \neq D_i(1)$ we can observe either one of its terms through appropriate choice of Z_i. We therefore focus on average causal effects in groups of people who can be induced to change treatments. Inferences about such average causal effects are made using changes in treatment status induced by treatment assignment, provided the assignment does affect the treatment.

At this point we introduce a compact notation to denote averages over the entire population or subpopulations. Let $E[g]$ denote the average over the population of N units of any function $g(\cdot)$ of $Z_i, D_i(1), D_i(0), Y_i(0,0), Y_i(0,1), Y_i(1,0)$, or $Y_i(1,1)$. Similarly, the average of $g(\cdot)$ over the subpopulation defined by some fixed value h_0 of some function $h(\cdot)$ will be denoted by $E[g|h(\cdot) = h_0]$. Finally, the relative size of the subpopulation satisfying $h(\cdot) = h_0$ is written as $P[h(\cdot) = h_0] = E[1_{h(\cdot)=h_0}]$, where $1_{\{\cdot\}}$ is the indicator function. We emphasize that this notation simply reflects averages and frequencies in a finite population or subpopulation.

The next assumption requires Z to have some effect on the average probability of treatment.

Assumption 4: Nonzero Average Causal Effect of Z on D.
The average causal effect of Z on $D, E[D_i(1) - D_i(0)]$ is not equal to zero.

The final assumption that we make, originally formulated by Imbens and Angrist (1994), says that there is no one who does the opposite of his assignment, no matter what the assignment.

Assumption 5: Monotonicity (Imbens and Angrist 1994).
$D_i(1) \geq D_i(0)$ for all $i = 1, \ldots, N$.

We refer to the combination of Assumptions 4 and 5, implying that $D_i(1) \geq D_i(0)$ with inequality for at least one unit as strong monotonicity.

Assumptions 1–5 lead to our formal definition of an instrument in the RCM.

Definition 3: Instrumental Variable for the Causal Effect of D on Y.

A variable Z is an instrumental variable for the causal effect of D on Y if: its average effect on D is nonzero, it satisfies

Table 1. Causal Effect of Z on Y, $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$, for the Population of Units Classified by $D_i(0)$ and $D_i(1)$

		$D_i(0)$	
		0	*1*
$D_i(1)$	0	$Y_i(1, 0) - Y_i(0, 0) = 0$ Never-taker	$Y_i(1, 0) - Y_i(0, 1) = -(Y_i(1) - Y_i(0))$ Defier
	1	$Y_i(1, 1) - Y_i(0, 0) = Y_i(1) - Y_i(0)$ Complier	$Y_i(1, 1) - Y_i(0, 1) = 0$ Always-taker

the exclusion restriction and the monotonicity assumption, it is randomly (or ignorably) assigned, and SUTVA holds (i.e., if Assumptions 1–5 hold).

3.3 Interpreting the Instrumental Variables Estimand

SUTVA and the exclusion restriction are sufficient to establish a fundamental relationship between the intention-to-treat effects of Z on Y and D and the causal effect of D on Y at the unit level:

$$\begin{aligned} &Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \\ &\quad = Y_i(D_i(1)) - Y_i(D_i(0)) \\ &\quad = [Y_i(1) \cdot D_i(1) + Y_i(0) \cdot (1 - D_i(1))] \\ &\qquad - [Y_i(1) \cdot D_i(0) + Y_i(0) \cdot (1 - D_i(0))] \\ &\quad = (Y_i(1) - Y_i(0)) \cdot (D_i(1) - D_i(0)). \qquad (9) \end{aligned}$$

Thus the causal effect of Z on Y for person i is the product of (i) the causal effect of D on Y and (ii) the causal effect of Z on D. We can therefore write the average causal effect of Z on Y as the weighted sum of average causal effects for two subpopulations, both with $D_i(0) \neq D_i(1)$:

$$\begin{aligned} &E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \\ &\quad = E[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0))] \\ &\quad = E[(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1] \\ &\qquad\qquad \cdot P[D_i(1) - D_i(0) = 1] \\ &\qquad - E[(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = -1] \\ &\qquad\qquad \cdot P[D_i(1) - D_i(0) = -1]. \qquad (10) \end{aligned}$$

The weights do not sum to 1 but rather to $P[D_i(0) \neq D_i(1)]$.

Equation (10) does not use monotonicity. The monotonicity assumption requires that $D_i(1) - D_i(0)$ equals either zero or one, so that the average causal effect of Z on Y equals the product of the average causal effect of D on Y for persons with $D_i(0) = 0$ and $D_i(1) = 1$ and their proportion in the population:

$$\begin{aligned} &E[Y_i(D_i(1), 1) - Y_i(D_i(0), 0)] \\ &\quad = E[(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1] \\ &\qquad\qquad \cdot P[D_i(1) - D_i(0) = 1]. \qquad (11) \end{aligned}$$

This establishes the relationship between the IV estimand and the causal effect of D on Y, which we summarize as a formal proposition.

Proposition 1: Causal Interpretation of the IV Estimand. Given Assumptions 1, 3, 4, and 5, the instrumental variables estimand is

$$\begin{aligned} &\frac{E[Y_i(D_i(1), 1) - Y_i(D_i(0), 0)]}{E[D_i(1) - D_i(0)]} \\ &\qquad = E[(Y_i(1) - Y_i(0))|D_i(1) - D_i(0) = 1]. \quad (12) \end{aligned}$$

We call this the Local Average Treatment Effect (LATE). This result follows directly from (11) combined with two facts: first, that the monotonicity assumption implies that $E[D_i(1) - D_i(0)]$ equals $P[D_i(1) - D_i(0) = 1]$, and second, that $E[D_i(1) - D_i(0]$ differs from zero.

Table 1 helps interpret this result. The four values of $(D_i(0), D_i(1))$ in this two-by-two table generate three distinct values of $D_i(1) - D_i(0)$. Individuals with $D_i(1) - D_i(0) = 1$ (bottom left) are induced to take the treatment by assignment to the treatment, and the causal effect of Z on Y is $Y_i(1) - Y_i(0)$ for individuals of this type, whom we refer to as *compliers.* A value of $D_i(1) - D_i(0) = 0$ (diagonal elements) implies that individual i does not change treatment status with the assigned treatment; the causal effect of Z on Y is zero for such individuals by the exclusion restriction. If $D_i(0) = D_i(1) = 0$, the individual is referred to as a *never-taker,* or in our application, a draft avoider; whereas if $D_i(0) = D_i(1) = 1$, the individual is an *always-taker,* or, in our application, a volunteer. Finally, individuals with $D_i(1) - D_i(0) = -1$ (top right) do the opposite of their assignment; they are induced to avoid the treatment by assignment to it, and induced to take the treatment by assignment to the control group. We call such individuals *defiers,* as suggested by Balke and Pearl (1993) in a comment on an earlier version of this paper (Angrist, Imbens, and Rubin 1993). The causal effect of Z on Y for these individuals is $Y_i(0) - Y_i(1)$. Finally, we refer to never-takers, always-takers, and defiers jointly as *noncompliers.* Note that these labels—compliers, defiers, never-takers, always-takers, and noncompliers—are simply definitions given SUTVA in this experiment and are not assumptions about individual behavior.

By virtue of the exclusion restriction, the two subpopulations corresponding to the two diagonal elements of Table 1 are characterized by a zero causal effect of Z on Y. By virtue of the monotonicity assumption there are no defiers, and the group corresponding to the top-right element in the table is empty. Finally, by virtue of Assumption 4, the proportion of the population in the cell corresponding to compliers differs from zero and is equal to the average causal effect of Z on D. Combined, these assumptions imply that the average causal effect of Z on Y is proportional to the average causal effect of D on Y for compliers. This is the result in Proposition 1.

Because we can estimate the two intention-to-treat estimands by virtue of random assignment, we can also estimate their ratio; that is, the IV estimand. The ratio of the usual unbiased estimators for the intention-to-treat estimands given in (7) and (8) is equal to the standard instrumental variables estimator for binary instruments given in (6). This estimator does not exploit all the implications of the model developed in this section. In Imbens and Rubin (1994a,b) we discuss implications of this model for estimation.

Finally, it is important to note that (under our assumptions) we cannot generally identify the specific members of the group of compliers, defined by $D_i(0) = 0, D_i(1) = 1$, for whom we can identify the average treatment effect. Thus, the local average treatment effect (i.e., the average causal effect for compliers) is not the average treatment effect for either the entire population or for a subpopulation identifiable from observed values. Stronger assumptions are needed for the identification of average causal effects for subpopulations identifiable from observed data. One assumption that achieves this is random assignment to a control group denied treatment, so that $D_i(0) = 0$ for all i (Zelen 1979; Angrist and Imbens 1991). For examples of other such assumptions see Heckman (1990), Robins and Tsiatis (1991), Efron and Feldman (1991), and Manski (1994).

4. COMPARING THE STRUCTURAL EQUATION AND POTENTIAL OUTCOMES FRAMEWORK

In Section 2 we described a structural equation model for the effect of military service on a health outcome using an indicator of draft eligibility as an instrument. Here we contrast that framework with the approach developed for the same problem in Section 3. In particular, we compare the formulation and clarity of the assumptions in each case. This comparison is useful because several authors have attributed the absence of structural equation methods in statistics to the manner in which such models are commonly formulated. For example, in his discussion of the connection between structural equation methods and path analysis, Goldberger (1972) quoted Moran (1961): "The main reason why Sewall Wright's method of path coefficients is often found difficult to understand is that expositions of the theory do not make clear what assumptions are made" (p. 988). Similarly, Holland (1988) writes, "it is not always evident how to verify assumptions made about [regression disturbances]. For example, why should [they] be independent of [Z] ... when the very definition of [the disturbances] involves [Z]" (p. 458).

4.1 The Exclusion Restriction and Ignorable Treatment Assignment

The econometric version of these assumptions requires that the disturbances in the response equation (1) and the participation equation (2) be uncorrelated with, or independent of, the assignment Z. In Imbens and Angrist (1994) this assumption is formulated in a framework using potential outcomes indexed only against the level of the treatment D. The framework we develop here separates this requirement into two assumptions about potentially observable quantities: the exclusion restriction, which says nothing about the treatment assignment mechanism, and ignorable treatment assignment, which says nothing about possible direct effects of assignment.

First, the exclusion restriction requires that the instrument have no effect on the outcome except through D. Thus to verify this assumption, the researcher must consider, at the unit level, the effect of changing the value of the instrument while holding the value·of the treatment fixed. To clarify the distinction between this formulation and the econometric formulation, consider the four subpopulations defined by the values of $D_i(0)$ and $D_i(1)$ in Table 1. Someone with $D_i(0) = D_i(1) = 1$ would always serve in the military with a low or high draft lottery number. It seems reasonable to assume that for such a person, the draft lottery number has no effect on health outcome. Next, consider someone with $D_i(0) = D_i(1) = 0$, who would have managed to avoid military service with a high or low lottery number. For someone exempted from military service for medical reasons, it seems plausible that there was no effect of the draft lottery number. But a draftee who managed to avoid military service by staying in school or moving abroad could experience an effect of Z on future life outcomes that would violate the exclusion restriction. For both these groups of noncompliers, the exclusion restriction requires the researcher to consider a difference in outcomes that were potentially observable, even though after the population was randomly allocated to treatment and control groups, only one of the outcomes was actually observed. In fact, if one could identify compliers and noncompliers, then it would be possible to test the exclusion restriction by comparing average outcomes for noncompliers by assignment status.

For compliers with $D_i(0) = 0, D_i(1) = 1$, the exclusion restriction compares outcomes that cannot be observed: it requires that $Y_i(0, D_i(0)) = Y_i(1, D_i(0))$ and $Y_i(1, D_i(1)) = Y_i(0, D_i(1))$. For this group, the exclusion restriction amounts to attributing the effect of Z on Y to the change in the treatment received D rather than to the change in assignment Z. Such an assumption is not innocuous, and efforts to ensure it form the rationale for blinding, double blinding, and using placebos in clinical trials. Nevertheless, it underlies most experimental evaluations in economics where blinding and placebos are impossible, and is often thought to be reasonable in those cases.

The second element embedded in the assumption of zero correlation between instruments and disturbances in the standard econometric formulation is that of random, or at least ignorable, treatment assignment Z. This assumption is trivially satisfied if physical randomization took place, as in the application in Section 6 where Z is a function of a lottery number. Our formulation makes clear that randomization of the instrument, though sufficient to allow unbiased estimation of the average treatment effect of Z on Y and of the average treatment effect of Z on D, does not imply that the IV estimand is interpretable as an average causal effect. In most applications of IV, however, the instrument is not randomly assigned, and this assumption must be argued more carefully. Examples include Angrist and Krueger's (1991) use of quarter of birth as an instrument for the effect of schooling on earnings, Card's (1993) use of distance to college as an instrument for the effect of schooling on earnings, and McClellan and Newhouse's (1994) use of relative distance to hospital as an instrument for the effect of catherization on mortality after acute myocardial infarction.

Whereas the exclusion restriction requires the researcher to contemplate the effect of specific treatments on outcomes, the ignorability assumption requires consideration of the assignment mechanism. Violations of these different assumptions can have different sources and consequences. In our view, pooling these assumptions into the single assumption of zero correlation between instruments and disturbances has led to confusion about the essence of the identifying assumptions and hinders assessment and communication of the plausibility of the underlying model.

4.2 The Monotonicity Condition

The monotonicity assumption rules out the existence of defiers, characterized by $D_i(0) = 1$ and $D_i(1) = 0$. Permutt and Hebel (1989) informally discussed a variant of this assumption in a reanalysis of a program designed to induce pregnant women to stop smoking. In that context, the assumption implies that everyone who would stop smoking if they were in the control group, which received no encouragement to stop smoking, would also stop smoking if encouraged to do so by being in the treatment group. Robins (1989) discussed the effect of this assumption on bounds on population average treatment effects. Monotonicity is implied by designs where those assigned to the control group are prevented from receiving the treatment, as in Zelen's (1979) single-consent designs.

Monotonicity has no explicit counterpart in the econometric formulation, but is implicit in the use of an equation with constant parameters for the relation between Z_i and D_i. The model developed in Section 3 suggests that the constant parameter assumption embodied in (2) is much stronger than needed. On the other hand, it is not sufficient to postulate a nonzero covariance between treatment and assignment, as in (5), for the interpretation of the IV estimand as an average of causal effects.

4.3 Reduced Form and Structural Parameters

Reduced-form parameters for the draft lottery application are the coefficients from a regression of Y on Z and D on Z. In our formulation, these are the average intention-to-treat effects under Assumptions 1 and 2. The structural parameter (β_1) is the average effect of the treatment itself on Y for the subpopulation that complies with assignment. The econometric approach does not distinguish between an effect for the entire population and an effect for the subpopulation of compliers. In our view LATE is structural in the Goldberger (1972) sense of representing a causal link, but not necessarily structural in the sense of representing a parameter that is invariant across populations. Despite this potential lack of generalizability, we view LATE as interesting (perhaps in combination with the intention-to-treat estimand) because it is an average of unit level causal effects of the treatment of interest. For example, for a potential recruit, the average effect of actual military service for a specific subpopulation is likely to be of greater interest than the population average effect of draft eligibility.

A similar rationale applies to clinical trials, which are often based on populations that are more homogeneous than, and not representative of, the population that will eventually be subjected to the treatment. The presumption in such cases, and in our analysis, is the average over the subpopulation of those whose behavior can be modified by assignment are likely to be informative about population averages of those who comply in the future, even if there is substantial heterogeneity in individual-level causal effects.

It should be stressed, however, that the assumptions needed for a causal interpretation of the instrumental variables estimand (Assumptions 1 and 3–5) are substantially stronger than those needed for the causal interpretation of the intention-to-treat estimand (Assumption 1). The plausibility of the additional assumptions (i.e., the exclusion restriction and the monotonicity assumption) must be taken into account when facing the choice to report estimates of the intention-to-treat estimands, of the IV estimands, or both.

5. SENSITIVITY OF THE IV ESTIMAND TO CRITICAL ASSUMPTIONS

The assumptions laid out in Section 3 are sufficient conditions for the identification of a meaningful average treatment effect. In this section we discuss the sensitivity of the IV estimand to deviations from the IV assumptions. As this discussion makes clear, violations of these assumptions need not be catastrophic. We focus on Assumptions 3 and 5 because they form the core of the IV approach. Assumption 4 (a nonzero average causal effect of Z on D) is conceptually straightforward and easy to check. Assumptions 1 and 2 are standard in the RCM approach, and sensitivity to particular violations of those assumptions has been previously discussed (e.g., Rosenbaum and Rubin 1983). In general the IV estimand is most likely to be sensitive to violations of the exclusion restriction and the monotonicity assumption when there are few compliers. In Section 6 we illustrate how this sensitivity analysis can be applied.

5.1 Violations of the Exclusion Restriction

First, we consider violations of the exclusion restriction, while maintaining the other assumptions, stability, and strong monotonicity. If subject i is a noncomplier, that is, $D_i(0) = D_i(1)$, then the causal effect of Z on Y is

$$H_i = Y_i(1, d) - Y_i(0, d), \tag{13}$$

where $d = 0$ if subject i is a never-taker and $d = 1$ if subject i is an always-taker. Under the exclusion restriction, $H_i = 0$ for all noncompliers.

Proposition 2. Given stability and strong monotonicity, but without the exclusion restriction for noncompliers, the IV estimand equals the Local Average Treatment Effect plus a bias term given by (14):

$$\begin{aligned} &\frac{E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{E[D_i(1) - D_i(0)]} \\ &\quad - E[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) | i \text{ is a complier}] \\ &= E[H_i | i \text{ is a noncomplier}] \cdot \frac{P[i \text{ is a noncomplier}]}{P[i \text{ is a complier}]} . \end{aligned} \tag{14}$$

The bias of the IV estimand relative to the Local Average Treatment Effect equals the average direct effect of Z on Y for noncompliers multiplied by the odds of being a noncomplier.

When there is a direct effect of assignment on the outcome for noncompliers, it is plausible that there is also a direct effect of assignment on outcome for compliers. Suppose that for each complier, assignment and treatment had additive effects on the outcome Y; that is,

$$Y_i(1, 0) - Y_i(0, 0) = Y_i(1, 1) - Y_i(0, 1),$$

for all compliers. Additivity for compliers allows us to define the causal effect of Z on Y for compliers as $H_i = Y_i(1, d) - Y_i(0, d)$ for $d = 0, 1$ [analogous to the definition for noncompliers given in (13)] and the causal effect of D on Y as $G_i = Y_i(z, 1) - Y_i(z, 0)$. We can then write the IV estimand as

$$\begin{aligned} &\frac{E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{E[D_i(1) - D_i(0)]} \\ &= E[G_i | i \text{ is a complier}] \\ &\quad + \frac{E[H_i]}{P[i \text{ is a complier}]} . \end{aligned} \tag{15}$$

The bias relative to the average causal effect of D on Y for compliers, the second term in (15), can also be written as

$$\begin{aligned} &E[H_i | i \text{ is a complier}] + E[H_i | i \text{ is a noncomplier}] \\ &\quad \cdot \frac{P[i \text{ is a noncomplier}]}{P[i \text{ is a complier}]} . \end{aligned} \tag{16}$$

The first term in the bias in (16) has nothing to do with noncompliance, but is the bias due to the direct effect of assignment for those who take the treatment. If compliance were perfect, the second term would be zero but the first term of the bias would still be present. The increased bias in the IV estimand due to noncompliance is directly proportional to the product of the average size of the direct effect of Z for noncompliers and the odds of noncompliance given monotonicity. The higher the correlation between the instrument and the treatment status (i.e., the "stronger" the instrument), the smaller the odds of noncompliance, and consequently the less sensitive the IV estimand is to violations of the exclusion assumption.

5.2 Violations of the Monotonicity Condition

Next we consider violations of the monotonicity assumption. Because we maintain the exclusion restriction, the causal effect of D on Y for person i with $D_i(1) \neq D_i(0)$ is still uniquely defined, and equal to $Y_i(1) - Y_i(0)$.

Proposition 3. Given stability, the exclusion restriction, and a nonzero average causal effect of Z on D, but without the monotonicity assumption, the IV estimand equals the Local Average Treatment Effect plus a bias term given by (17):

$$\begin{aligned} &E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]/E[D_i(1) - D_i(0)] \\ &\quad - E[Y_i(1) - Y_i(0) | i \text{ is a complier}] \\ &= -\lambda \cdot \{E[Y_i(1) - Y_i(0) | i \text{ is a defier}] \\ &\quad - E[Y_i(1) - Y_i(0) | i \text{ is a complier}]\} , \end{aligned} \tag{17}$$

where

$$\lambda = \frac{P(i \text{ is a defier})}{P(i \text{ is a complier}) - P(i \text{ is a defier})} . \tag{18}$$

The bias due to violations of monotonicity is composed of two factors. The first factor, $\lambda = P(i \text{ is a defier})/(P(i \text{ is a complier}) - P(i \text{ is a defier}))$, is related to the proportion of defiers and is equal to zero under the monotonicity assumption. The smaller the proportion of defiers, the smaller the bias will be from violations of the monotonicity assumption. However, because the denominator of this factor is the average causal effect of Z on D, the bias can be large even if there are few defiers, as long as the average causal effect of Z on D is small. Note again that the stronger the instrument, the less sensitive the IV estimand is to violations of the monotonicity assumption. The second factor is the difference in average causal effects of D on Y for the compliers and defiers. If the average causal effects of D on Y are identical for defiers and compliers, violations of the monotonicity assumption generate no bias. The less variation there is in the causal effect of D on Y, the smaller the bias from violations of the monotonicity assumption.

Without monotonicity, the IV estimand can also be written as

$$\begin{aligned} &(1 + \lambda) \cdot E[Y_i(1) - Y_i(0) | i \text{ is a complier}] \\ &\qquad - \lambda \cdot E[Y_i(1) - Y_i(0) | i \text{ is a defier}], \end{aligned}$$

with λ as defined in (18). In this representation, the estimand is still a weighted average of average treatment effects despite the violation of the monotonicity assumption, but the weights are always outside the unit interval because $\lambda > 0$.

6. AN APPLICATION: THE EFFECT OF MILITARY SERVICE ON CIVILIAN MORTALITY

Hearst, Newman, and Hulley (1986) showed that men with low lottery numbers in the Vietnam Era draft lottery (i.e., men with $Z_i = 1$) had elevated mortality after their discharge from the military. The authors attribute this elevated mortality to the detrimental effect of serving in the military during wartime on well-being. Similarly, Angrist (1990) attributed differences in subsequent earnings by lottery number to the effect of serving in the military on earnings. These conclusions are primarily based on the fact that between 1970 and 1973, priority for the draft was randomly assigned in a lottery using dates of birth. Each date of birth in the cohorts at risk of being drafted was assigned a *random sequence number* (RSN) from 1–365. The Selective Service called men for induction by RSN up to a ceiling determined by the defense department. Men born in 1950 were potentially drafted up to RSN 195 in 1970, men born in 1951 were potentially drafted up to RSN 125 in 1971, and men born in 1952 were potentially drafted up to RSN 95 in 1972. We refer the reader to Hearst et al. (1986) and Angrist (1990) for further details on these data and the draft.

In their paper, Hearst et al. focused on the difference in mortality risk by draft status. For example, they compare the number of deaths of men born in 1950 with RSN below 195 to the number of deaths of men born in 1950 with RSN above 195. Our purpose in returning to this example is twofold. First, we discuss the validity of Assumptions 1–5 in this context. Second, we show how the sensitivity of the estimated average treatment effect to violations of the exclusion restriction and the monotonicity assumption can be explored using the results from the previous section.

6.1 Assessment of Assumptions 1–5

The potential outcome in this example, $Y_i(z, d)$, is an indicator variable equal to one if person i would have died between 1974 and 1983 given lottery assignment z and military service indicator d. To distinguish this from mortality during the war period, we refer to Y_i as civilian mortality. For simplicity, we ignore the effect that mortality during the war might have on the size of the population at risk.

For a valid causal interpretation of the IV estimand, we require:

- SUTVA, Assumption 1: The veteran status of any man at risk of being drafted in the lottery was not affected by the draft status of others at risk of being drafted, and, similarly, that the civilian mortality of any such man was not affected by the draft status of others;
- Ignorable Assignment, Assumption 2: Assignment of draft status was random;
- Exclusion restriction, Assumption 3: Civilian mortality risk was not affected by draft status once veteran status is taken into account;
- Nonzero Average Causal Effect of Z on D, Assumption 4: Having a low lottery number increases the average probability of service;
- Monotonicity assumption, Assumption 5: There is no one who would have served if given a high lottery number, but not if given a low lottery number.

Although we believe these assumptions are plausible, a case can be made for violations of most. For example, it has been argued that the fraction of a cohort that served in the military affects the civilian labor market response to veterans (De Tray 1982). If this assertion is true, then the SUTVA assumption very likely does not hold. Another reason for possible violations of SUTVA is that people not drafted may be induced to serve in the military by friends who were drafted.

There is also some evidence that some men with low lottery numbers changed their educational plans so as to retain draft deferments and avoid the conscription (Angrist and Krueger 1992b). If so, then the exclusion restriction could be violated, because draft status may have affected civilian outcomes through channels other than veteran status. We return to this issue in some detail shortly.

Monotonicity would be violated if, for example, someone, who would have volunteered for the Navy when not at risk of being drafted because of a high lottery number, would have chosen to avoid military service altogether when at risk of being drafted because of a low lottery number. It seems unlikely that there were many in the population in this category.

It is clear that the Assumption 4 is satisfied because the likelihood of serving in the military sharply increases with draft status.

Another uncontroversial assumption is the ignorability of treatment assignment, which allows simple unbiased estimation of the average causal effects of Z on D and of Z on Y. Although there is some evidence that the first lottery, which was executed using a poorly designed physical randomization, was not actually random (Fienberg 1971) it nevertheless is almost certainly ignorable. Ignoring this complication and postponing consideration of the possible problems with the exclusion restriction and the monotonicity condition, we forge ahead with the IV approach.

6.2 The Instrumental Variables Estimates

Table 2 presents data and some estimates of the effects of military service on civilian mortality for white men born in 1950 and 1951 by year of birth and draft status. Column 3 shows the number of deaths in both Pennsylvania and California between 1974–1983. Columns 5 and 6 show the average number of civilian deaths and suicides respectively per 1,000, computed as the number of deaths divided by the population at risk estimated using the 1970 census. Column 7 shows the frequency of veteran status, estimated from the 1984 Survey of Income and Program Participation (SIPP). In columns 5–7, the entries in the third pair of rows give the difference in probability of death, suicide, and veteran status between those with low and high lottery numbers (draft eligible or not). The fourth pair of rows in columns 5 and 6 give the ratio of these differences to the difference in the probability of being veteran by draft eligibility. These are the standard IV estimates. An alternative approach to

Table 2. Data on Civilian Mortality for White Men Born in 1950 and 1951

Year	Draft eligibility[a]	Number of deaths[b]	Number of suicides[c]	Probability of death[d]	Probability of suicide	Probability of military service[e]
1950	Yes	2,601	436	.0204	.0034	.3527
				(.0004)	(.0002)	(.0325)
	No	2,169	352	.0195	.0032	.1934
				(.0004)	(.0002)	(.0233)
Difference (Yes minus No)				.0009	.0002	.1593
				(.0006)	(.0002)	(.0401)
IV estimates[f]				.0056	.0013	
				(.0040)	(.0013)	
1951	Yes	1,494	279	.0170	.0032	.2831
				(.0004)	(.0002)	(.0390)
	No	2,823	480	.0168	.0029	.1468
				(.0003)	(.0001)	(.0180)
Difference (Yes minus No)				.0002	.0003	.1362
				(.0005)	(.0002)	(.0429)
IV estimates				.0015	.0022	
				(.0037)	(.0016)	

[a] Determined by lottery number cutoff: RSN 195 for men born in 1950, and RSN 125 for men born in 1951.

[b] From California and Pennsylvania administrative records, all deaths 1974–1983. Data sources and methods documented by Hearst et al. (1986). Note: Sample sizes differ from Hearst et al., because non-U.S.-born are included to match SIPP data in the last column.

[c] The mortality figures are tabulated from the data set analyzed by Hearst et al. (1986).

[d] The estimated population at risk is from the author's tabulation of 1970 census data. Estimates by draft-eligibility status are computed assuming a uniform distribution of lottery numbers. Standard errors are given in parentheses.

[e] These figures are taken from Angrist (1990), table 2, and were tabulated using a special version of the SIPP that has been matched to indicators of draft eligibility. Note that probabilities estimated using the SIPP are for the entire country and do not take account of morality. The impact of mortality on differences in the probability of being a veteran by eligibility status is small enough to have only trivial consequences for the estimation.

[f] The standard errors, following econometric practice (e.g. Imbens and Angrist 1994), were calculated based on a normal approximation to the sampling distribution of the ratio of the difference in estimated probability of death/suicide and the difference in estimated probability of serving. We assume independence of numerator and denominator because they were calculated from different data sets. Pooled estimates show a statistically significant increase in risk at conventional significance levels (e.g., Hearst, Newman, Hulley 1986).

estimating the local average treatment effect, which takes into account the full implications of the assumptions, is provided in Imbens and Rubin (1994b).

As a specific example, consider men born in 1950. Of the men with low lottery numbers ($Z_i = 1$), 35.3% actually served in the military. Of those who had high lottery numbers ($Z_i = 0$), only 19.3% served in the military. Random assignment of draft status suggests that draft status had a causal effect that increased the probability of serving by an estimated 15.9% on average. Similarly, of those with low lottery numbers, 2.04% died between 1974 and 1983, compared to 1.95% of those who had high lottery numbers. The difference of .09% can be interpreted as an estimate of the average causal effect of draft status on civilian mortality. Assuming that these estimated causal effects are population averages, the ratio of these two causal effects of draft status is, under the Assumptions 1–5, the causal effect of military service on civilian mortality for the 15.9% who were induced by the draft to serve in the military. For this group, the average causal effect is .56%, which amounts to approximately a 25% increase in the probability of death (given average mortality rates around 7%). These estimates highlight the fact that the IV estimator does not require observations on individuals; sample averages of outcomes and treatment indicators by values of the instruments are sufficient. In applications like the one discussed here, these moments are drawn from different data sets. (For a detailed discussion of IV estimation with moments from two data sets, see Angrist and Krueger 1992a.)

6.3 Sensitivity to the Exclusion Restriction

Suppose that the exclusion restriction is violated because men with low lottery numbers were more likely to stay in school. A schooling–lottery connection could arise because, for much of the Vietnam period, college and graduate students were exempt from the draft. Although new graduate student deferments were eliminated in 1967 and new undergraduate deferments were eliminated in December 1971, many of the men with low lottery numbers in 1970 and 1971 could have postponed conscription by staying in school. Working with special versions of the March 1979 and March 1981–1985 Current Population Surveys (CPS's), Angrist and Krueger (1992b) showed that men born in 1951 with lottery numbers 1–75 had completed .358 more years of schooling than men with lottery numbers above 150, who were not drafted.

How much bias in estimates of the effect of military service on mortality is this correlation between lottery numbers and schooling likely to generate? Addressing this question requires data on the connection between schooling and mortality. The relationship between socioeconomic variables and mortality is uncertain and the subject of considerable research in epidemiology and social science. (An early study in this area is Kitagawa and Hauser 1973.) For the purposes of illustration, we have taken estimates from Duleep's (1986) study of socioeconomic variables and mortality using men surveyed in the March 1973 CPS and linked to 1973–1978 Social Security data. Estimates presented in Table 1 of Duleep (1986) suggest that married white men 25 years old with 1–3 years of college have mortality rates roughly .0017 per thousand higher than do men with only high school degrees.

Assume that the excess mortality among men with some college accumulates linearly, so that an additional year of schooling raises mortality by .0017 × (1/3) = .00056. Men with low lottery numbers may have as much as .358 more years of schooling than men with high lottery numbers. Thus an estimate of the mortality difference attributable to the effect of draft status on schooling is .358 × .00056 = .00019, essentially as large as the .0002 observed difference in mortality by draft status for white men born in 1951. Assuming additive causal effects of education and military service on mortality, the bias formula (15) applied to this example is $E[H_i]/(E[D_i(1) - D_i(0)])$, which is estimated by $.00019/.1362 = .0014$ because there is a .1362 difference in the probability being a veteran by draft eligibility status. Thus taking account of this potential bias could eliminate the estimated .0015 impact of veteran status on civilian mortality!

This calculation illustrates the cautions that should accompany the IV estimates. But the extent to which the causal interpretation of the estimates in Table 2 should be discounted in light of these findings is unclear. First, there is no evidence of a schooling–lottery number connection for the 1950 cohort, yet lottery-based estimates of the effects of service are even larger for men born in 1950 than for the 1951 cohort used in the illustration. Second, the schooling–mortality connection is not well determined [the Duleep (1986) estimate used here is not actually significantly different from zero], and this relationship is also subject to sign reversals. For example, although men with some college have higher mortality than high school–only graduates, the Duleep study showed almost no difference between the mortality of high school only graduates and college graduates. Thus, a calculation based solely on graduates would indicate no bias.

6.4 Sensitivity to the Monotonicity Assumption

Without monotonicity, the average causal effect of Z on D estimates the difference between the proportions of compliers and defiers. Table 2 therefore suggests that 15.93% more people are compliers than defiers. Suppose that 5% of the population are defiers. This would imply that about 21% of the population are compliers, and that the multiplier $P[i \text{ is a defier}]/(P[i \text{ is a complier}] - P[i \text{ is a defier}])$ could be as large as .33 rather than zero, as required by monotonicity. Next, suppose that we assume the difference between average treatment effects for compliers and defiers is at most .0041. This number was chosen because the range of IV estimates in Table 2 (.0056 for 1950 and .0015 for 1951) is equal to this amount. This implies that the estimated average treatment effect for compliers could be as small as $.0056 - .33 \times .0041 = .0042$ or as large as $.0056 + .33 \times .0041 = .0070$. To reverse the sign of the average causal effect through violations of the monotonicity assumption would therefore require the presence of an implausibly large group of defiers, or very large differences between average effects for compliers and defiers.

7. CONCLUSION

In this article we have outlined a framework for causal inference in settings where random assignment has taken place, but compliance is not perfect; that is, the treatment received is nonignorable. In an attempt to estimate the effect of receipt of treatment, rather than assignment of treatment as in intention-to-treat analysis, we make use of instrumental variables. This approach has long been used by economists in the context of regression models with constant treatment effects. We show that this technique can be fit into the Rubin Causal Model and used for causal inference without assuming constant treatment effects. The advantages of embedding this approach in the RCM are twofold. First, it makes the nature of the identifying assumptions more transparent. Second, it allows us to consider the sensitivity of results to deviations from these assumptions in a straightforward manner. We hope that the approach outlined in this article serves to make the IV approach more accessible to statisticians, while helping economists understand and interpret the strong assumptions required for a causal interpretation of IV estimates.

[Received June 1993. Revised December 1995.]

REFERENCES

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

Angrist, J. D., and Imbens, G. W. (1991), "Sources of Identifying Information in Evaluation Models," Technical Working Paper 117, National Bureau of Economic Research.

——— (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1993), "Identification of Causal Effects Using Instrumental Variables," Technical Working Paper 136, National Bureau of Economic Research.

Angrist, J., Graddy, K., and Imbens, G. W. (1995), "Nonparametric Demand Analysis With an Application to the Demand for Fresh Fish," Technical Working Paper 178 NBER April '95, Massachusetts Institute of Technology.

Angrist, J. D., and Krueger, A. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 979–1014.

——— (1992a), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables With Moments From Two Samples," *Journal of the American Statistical Association*, 87, 328–336.

——— (1992b), "Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery," Working Paper 4067, National Bureau of Economic Research.

Balke, A., and Pearl, J. (1993), "Nonparametric Bounds on Causal Effects From Partial Compliance Data," Technical Report R-199, University of California, Los Angeles, Computer Science Dept.

Bowden, R. J., and Turkington, D. A. (1984), *Instrumental Variables*, Cambridge, U.K.: Cambridge University Press.

Card, D. (1993), "Using Geographic Variation in College Proximity to Estimate the Returns to Schooling," Working Paper 4483, National Bureau of Economic Research.

Cox, D. R. (1958), *Planning of Experiments*, New York: John Wiley.

De Tray, D. (1982), "Veteran Status as a Screening Device," *American Economic Review*, 72, 133–142.

Duleep, H. O. (1986), "Measuring the Effect of Income on Adult Mortality Using Longitudinal Administrative Record Data," *Journal of Human Resources*, 21, 238–251.

Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 22, 23–32.

Efron, B., and Feldman, D. (1991), "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association*, 86, 9–26.

Fienberg, S. (1971), "Randomization and Social Affairs: The 1970 Draft Lottery," *Science*, 171, 255–261.

Fisher, R. A. (1918), "The Causes of Human Variability," *Eugenics Review*, 10, 213–220.

——— (1925), *Statistical Methods for Research Workers*, London: Oliver & Boyd.

Goldberger, A. S. (1972), "Structural Equation Methods in the Social Sciences," *Econometrica*, 40, 979–1001.

Greenland, S., and Robins, J. M. (1986), "Identifiability, Exchangeability, and Epidemiological Confounding," *International Journal of Epidemiology*, 15, 413–419.

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

——— (1944), "The Probability Approach in Econometrics," *Econometrica*, 12 (Supplement), 1–115.

Hearst, N., Newman, T., and Hulley, S. (1986), "Delayed Effects of the Military Draft on Mortality: A Randomized Natural Experiment," *New England Journal of Medicine*, 314, 620–624.

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

Heckman, J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, eds. Heckman and Singer, New York: Cambridge University Press, pp. 156–245.

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

——— (1988), "Causal Inference, Path Analysis, and Recursive Structural Equations Models," (with discussion) in *Sociological Methodology*, ed. Washington, DC: American Sociological Association, pp. 449–493.

Imbens, G. W., and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Imbens, G. W., and Rubin, D. B. (1994a), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," Harvard University, Dept. of Economics.

——— (1994b), "On the Fragility of Instrumental Variables Estimators," Working Paper 1356, Harvard Institute of Economic Research.

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

Kitagawa, E. M., and Hauser, P. M. (1973), *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*, Cambridge, MA: Harvard University Press.

Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620.

Lee, Y., Ellenberg, J., Hirtz, D., and Nelson, K. (1991), "Analysis of Clinical Trials by Treatment Actually Received: Is It Really an Option?," *Statistics in Medicine*, 10, 1595–1605.

Little, R. (1985), "A Note About Models for Selectivity Bias," *Econometrica*, 53, 1469–1474.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, U.K.: Cambridge University Press.

Manski, C. F. (1994), "The Selection Problem," in *Advances in Econometrics*, ed. C. Sims, New York: Cambridge University Press, pp. 143–170.

McClellan, M., and Newhouse, J. (1994), "Marginal Benefits of Medical Treatment Intensity: Acute Myocardial Infarction in the Elderly," mimeo, National Bureau of Economic Research.

Moran, P. (1961), "Path Coefficients Reconsidered," *Australian Journal of Statistics*, 3, 87–93.

Morgan, M. (1990), *The History of Econometric Ideas*, Cambridge, U.K.: Cambridge University Press.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," translated in *Statistical Science*, 5, 465–480, 1990.

Permutt, T., and Hebel, J. (1989), "Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight," *Biometrics*, 45, 619–622.

Robins, J. M. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. by L. Sechrest, H. Freeman, and A. Bailey, Washington, DC: U.S. Public Health Service.

Robins, J. M., and Tsiatis, A. A. (1991), "Correcting for Non-Compliance in Randomized Trials Using Rank-Preserving Structural Failure Time Models," *Communications in Statistics, Part A—Theory and Methods*, 20, 2069–2631.

Rosenbaum, P., and Rubin, D. B. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

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

——— (1978), "Bayesian Inference for Causal Effects," *The Annals of Statistics*, 6, 34–58.

——— (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, 75, 591–593.

——— (1990), Comment: "Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.

——— (1991), "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism," *Biometrics*, 47, 1213–1234.

Schultz, H. (1928), *Statistical Laws of Supply and Demand*, Chicago: University of Chicago Press.

Sommer, A., and Zeger, S. (1991), 'On Estimating Efficacy from Clinical Trials," *Statistics in Medicine*, 10, 45–52.

Wright, S. (1928), Appendix to *The Tariff on Animal and Vegetable Oils*, by P. G. Wright, New York: MacMillan.

——— (1934), "The Method of Path Coefficients," *Annals of Mathematical Statistics*, 5, 161–215.

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

Comment

James M. ROBINS and Sander GREENLAND

We wish to complement the interesting paper by Angrist, Imbens, and Rubin (AIR) by offering several alternative analytic strategies. We focus on randomized drug treatment trials. In their discussion of noncompliance, AIR focuses on estimating the local average treatment effect (LATE), which is the average effect of treatment in the compliers. In contrast, Robins (1989) focused on estimation of the global average treatment effect (ATE) in the entire study population. Both LATE and ATE differ from the intent-to-treat (ITT) parameter, which is the average effect of treatment assignment. We show that in a typical placebo-controlled trial, all three parameters will equal zero under the sharp null hypothesis of no treatment effect. We argue that under the alternative, the ATE parameter can be of greater public health interest than the LATE or ITT parameter. We review results of Robins, Manski, and Balke and Pearl on the estimation of the ATE parameter. We show that in trials comparing a new therapy to a standard therapy, the null hypothesis of bioequivalence does not imply the ITT parameter is zero, and thus the ITT parameter is often of no public health interest. We review results on the estimation of the ATE parameter in bioequivalence trials.

Following AIR, for subject $i = 1, \dots, n, Z_i$ denotes the dichotomous randomization indicator (i.e. treatment arm); $D_i(z)$ denotes the actual treatment when randomized to arm $z, z = 0, 1$; $D_i = D_i(Z_i)$ denotes the observed treatment; $Y_i(z, d)$ denotes the outcome that would be observed if randomized to arm z and treatment d were taken, $z = 0, 1, d = 0, 1$; $Y_i = Y_i(Z_i, D_i(Z_i))$ denotes the observed outcome; and expectations are sample averages. This notation incorporates Rubin's stable unit treatment value assumption (SUTVA) assumption. Like AIR, we shall ignore sampling variability by restricting attention to estimands—the large-sample limits of estimators. The foregoing notation is sufficient to describe a trial in which each subject is either on or off a single active treatment; for example, a placebo-controlled trial. However, it is not sufficient to describe a bioequivalence trial that compares a new therapy to a standard therapy, because in a bioequivalence trial, noncompliers may choose to take no drug at all. Thus d needs to be at least trichotomous, corresponding to standard therapy, new therapy, and no therapy.

1. TRIALS WITH A SINGLE ACTIVE TREATMENT

Robins (1989) and AIR considered the analysis of randomized trials of a single active treatment with noncompliance under (1) the exclusion restriction $Y_i(1, d) = Y_i(0, d) \equiv Y_i(d)$, for all i and d, (2) the monotonicity assumption that $D_i(1) \geq D_i(0)$ for all i, and (3) the random assignment assumption that $D_i(z), Y_i(z, d), z = 0, 1, d = 0, 1$ are jointly independent of Z_i which we write as $\{D_i(z), Y_i(z, d); z = (0, 1), d = (0, 1)\} \coprod Z_i$. Robins and AIR also investigated the sensitivity of inferences to violations of these assumptions. (The three assumptions correspond exactly to assumptions (1)–(3) in Robins 1989, p. 123.) Robins studied the average treatment effect (ATE_z) controlling for treatment assignment z; that is, $E[Y_i(z, 1) - Y_i(z, 0)]$. Under the exclusion restriction, $\text{ATE}_1 = \text{ATE}_0 \equiv \text{ATE} \equiv E[Y_i(1) - Y_i(0)]$. AIR studied the local average treatment effect among the compliers, which is $E[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1]$ under the exclusion restriction. In contrast, the ITT parameter is $E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]$. Under random assignment, the ITT parameter equals the difference in treatment arm–specific means $E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)$.

Assuming the exclusion restriction, all three parameters will be zero under the sharp null hypothesis of no causal effect of D on Y; that is, $Y_i(1) = Y_i(0)$ for all i. Even with noncompliance, the large investment in randomized trials is considered worthwhile because, in contrast to a nonrandomized study, valid tests of the sharp null hypothesis can be obtained from the observed data by comparing treatment arm–specific means. Debates over which of the three parameters should represent the causal parameter of interest arise under alternatives to the sharp null.

A common argument in favor of the ITT parameter is that it corresponds to the overall treatment effect that would be realized if the treatment were actually adopted in the community. But this argument assumes that the noncompliance rate observed in the trial would equal the subsequent rate in the community, which may often not be the case. For example, once the treatment is proven to be efficacious in a trial, then nearly all individuals in the community may be willing to stringently comply with the treatment protocol (Robins 1989). In such a case, if the study subjects are representative of the community, then the ATE parameter, rather than the ITT or LATE parameter, would correspond to the public health parameter of interest. An advantage of the LATE parameter is that it is identifiable under monotonicity, whereas the ATE parameter is not. But unless no subject in the control arm takes active treatment, the subset of the study population for whom the LATE parameter is the treatment effect (i.e., the compliers) is itself nonidentifiable (AIR 1995). As discussed later, the LATE parameter is not identifiable under more complex noncompliance patterns, even if monotonicity holds.

The distribution of the observed data only determines bounds for the ATE parameter. Assuming Y_i dichotomous,

James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. Sander Greenland is Professor of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095.

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

Robins (1989) calculated bounds for ATE_z under the $2^3 = 8$ combinations of the truth or falsity of the exclusion restriction, monotonicity, and random assignment. Related results were independently obtained by Manski (1990, 1994). Interestingly, the bounds for the ATE_z parameter do not depend on the monotonicity assumption. Some argue against reporting bounds for nonidentifiable parameters, because bounds are often so wide as to be useless for making public health decisions. But we view the latter problem as a reason *for* reporting bounds in conjunction with other analyses: Wide bounds make clear that the degree to which public health decisions are dependent on merging the data with strong prior beliefs. Even when the ITT null hypothesis of equality of treatment arm–specific means is rejected, the bounds may appropriately include zero. If treatment benefits some subjects and harms others, the ATE parameter may be zero even though both the sharp and ITT null hypotheses are false. Conversely, the ATE parameter may be nonzero under the ITT null, seriously complicating the interpretation of tests of the ITT null in trials with substantial noncompliance. But there are times that bounds can be quite informative. For example, Balke and Pearl (1993) reanalyzed data from the Lipid Research Clinic's Coronary Primary Prevention Trial and showed that the Robins–Manski bounds are quite informative, with the lower bound lying far above the null value of zero.

Henceforth we assume that both the exclusion restriction and the random assignment assumption hold. Balke and Pearl (1993) showed that for certain distributions of the observed data, the Robins–Manski bounds, $-1 + \max_z\{\text{pr}(Y_i = 1, D_i = 1 | Z_i = z)\} + \max_z\{\text{pr}(Y_i = 0, D_i = 0 | Z_i = z)\} \le \text{ATE} \le 1 - \max_z\{\text{pr}(Y_i = 0, D_i = 1 | Z_i = z)\} - \max_z\{\text{pr}(Y_i = 1, D_i = 0 | Z_i = z)\}$, are not sharp. They derived narrower sharp bounds for these distributions. Specifically, when the bounds do not coincide, the Robins–Manski bounds are sharp under the weak randomization assumption $Y_i(d) \coprod Z_i, d = 0, 1$ (Manski 1994), whereas the Balke–Pearl bounds are sharp under the strong randomization assumption $\{Y_i(0), Y_i(1)\} \coprod Z_i$. The strong randomization assumption is satisfied in a truly randomized study. The even stronger randomization assumption $\{Y_i(0), Y_i(1), D_i(0), D_i(0)\} \coprod Z_i$, although appropriate in a randomized trial, does not change the ATE bounds. Balke and Pearl (1993) also showed that there are distributions of the observed data that are incompatible with jointly assuming the exclusion restriction and random assignment of Z.

If one wishes to identify the ATE parameter, further strong nonidentifiable assumptions must be added. Robins (1989, p. 122, assumptions 5–8) provided four different identifying assumptions and computed the ATE parameter under each as a form of sensitivity analysis. Assumption 6 of Robins (1989) implies that the ATE parameter equals the instrumental variable (IV) estimand $\{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)\}/\{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)\}$. This result assumes neither monotonicity nor that the subject-specific treatment effects $Y_i(1) - Y_i(0)$ are constant. Assumption 6 of Robins (1989) is the assumption that both (a) within each treatment arm z, the average treatment effect is the same for the treated $(D_i = 1)$ as for the untreated $(D_i = 0)$, i.e., $E\{Y_i(1) - Y_i(0) | Z_i = z, D_i = 1\} = E\{Y_i(1) - Y_i(0) | Z_i = z, D_i = 0\}$, and (b) the average treatment effect among the treated is the same for both treatment arms; that is, $E\{Y_i(1) - Y_i(0) | Z_i = 1, D_i = 1\} = E\{Y_i(1) - Y_i(0) | Z_i = 0, D_i = 1\}$. Assumptions (a) and (b) are always true under the sharp null. Robins (1989, sec. 16, 1994) introduced the class of structural nested mean models (SNMM's) for the average effect of treatment on the treated. Assumption (b) is equivalent to assuming a simple SNMM. Robins (1994) proved that the SNMM (b) alone implies that the average treatment effect in the treated $E\{Y_i(1) - Y_i(0) | D_i = 1\}$ is the IV estimand. The additional assumption (a) guarantees that the average treatment effect in the untreated $(D_i = 0)$ equals that in the treated $(D_i = 1)$, and hence that the ATE parameter equals the IV estimand. An estimand or parameter that is zero if and only if the ITT parameter is zero is called ITT null consistent. Because, when defined, the IV estimand is ITT null consistent, the ATE parameter is also ITT null consistent under assumptions (a) and (b).

2. BIOEQUIVALENCE TRIALS

A critical difference between a trial with a single active therapy and a bio equivalence trial is that in the presence of noncompliance, the sharp null hypothesis of the bioequivalence of the two therapies does not imply equality of treatment arm–specific mean outcomes. Consider a randomized bioequivalence trial in which a new therapy $(D = 1)$ is compared to standard proven therapy $(D = 0)$. Suppose that all subjects are initially compliant, but 50% of subjects assigned to standard therapy $(Z = 0)$ and 20% of subjects assigned to the new therapy $(Z = 1)$ later become noncompliant and stop all therapy due to mild, easily palliated side effects. Even if the ITT parameter $E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]$ demonstrated a beneficial effect of assignment to the new therapy, the benefit might be wholly attributable to the high noncompliance rate in the standard therapy arm. Thus the ITT test of equality of treatment arm–specific means may not be of regulatory or public health interest. To formalize our point, we consider four treatments: always remain on standard therapy $(D = 0)$; always remain on the new therapy $(D = 1)$; begin standard therapy, then stop all therapy $(D = 2)$; and begin the new therapy, then stop all therapy $(D = 3)$. If all therapy terminations were due to mild, easily palliated side effects, then the sharp bioequivalence null hypothesis of medical interest is $Y_i(1) = Y_i(0)$ for all i. But this null hypothesis does not imply the ITT null $E[Y_i | Z = 1] = E[Y_i | Z = 0]$. The ITT null is implied by the sharp null hypothesis that $Y_i(d) = Y_i$, for all i and $d = 0, 1, 2, 3$; this latter hypothesis is not of interest because it implies that being off therapy was as efficacious as being on the standard proven therapy, which is already known to be false.

In a bioequivalence trial, possible parameters of interest would be the ATE parameter $E[Y_i(1) - Y_i(0)]$ or the LATE parameter $E[Y_i(1) - Y_i(0)|D_i(1) = 1, D_i(0) = 0]$. Neither of these parameters is identifiable from bioequivalence trial data, even under a monotonicity assumption. Sharp bounds for the ATE parameter are $-1 + \text{pr}(Y_i = 1, D_i = 1|Z_i = 1) + \text{pr}(Y_i = 0, D_i = 0|Z_i = 0) \leq \text{ATE} \leq 1 - \text{pr}(Y_i = 0, D_i = 1|Z_i = 1) - \text{pr}(Y_i = 1, D_i = 0|Z_i = 0)$. If one wishes to identify the ATE parameter, then further strong nonidentifiable assumptions must be added. Heyting, Tolboom, and Essers (1992), Robins (1987), Robins and Rotnitzky (1992), and Robins, Rotnitzky, and Zhao (1995) have studied identifying assumptions for ATE in this setting. If the decision to quit therapy is essentially a second randomization (i.e., $Y_i(d) \coprod D_i|Z_i$ for $d = 0, 1$) we say that the noncompliance is random. In that case, $E[Y_i(d)]$ is identifiable and equal to $E[Y_i|Z_i = d, D_i = d]$ for $d = 0, 1$. If, as is usually the case, one does not believe compliance is random given Z_i, then one can try to collect data on additional pre- or post-randomization covariates L such that compliance is random conditional on the covariates; that is, $Y_i(d) \coprod D_i|Z_i, L_i, d = 0, 1$. Under this assumption, $E[Y_i(d)]$ is identifiable and, for discrete L, equals $\sum_l E[Y_i|Z_i = d, D_i = d, L_i = l]\text{pr}[L_i = l|Z_i = d]$ for $d = 0, 1$. Robins (1987) called this formula the G computation algorithm formula given covariates L. This formula can also be written as the inverse probability of censoring weighted (IPCW) estimand $E[Y_i I(Z_i = d, D_i = d)/\pi_{1d}\pi_{2d}(L_i)]$, where $\pi_{1d} = \text{pr}[Z_i = d]$ and $\pi_{2d}(L_i) \equiv \text{pr}[D_i = d|Z_i = d, L_i]$.

The IPCW estimand is easily generalized to allow investigation of the sensitivity of the estimate of $E[Y_i(d)]$ to the assumption $Y_i(d) \coprod D_i|Z_i, L_i$. Let $\pi_{2d}(L_i, y) = \text{pr}[D_i = d|Z_i = d, L_i, Y_i(d) = y]$ be the probability of treatment $D = d$ given $Z_i = d, L_i$, and $Y_i(d) = y$. Note that $\pi_{2d}(L_i, y)$ is not identifiable, because we do not observe $Y_i(d)$ for subjects for whom $D_i \neq d$. In a sensitivity analysis, we select plausible functions $\pi_{2d}(L_i, y)$ based on our prior beliefs. For a given $\pi_{2d}(L_i, y), E[Y_i(d)]$ is given by the IPCW estimand with $\pi_{2d}(L_i, Y_i)$ substituted for $\pi_{2d}(L_i)$ Robins et al. (1995, p. 118) also considered estimation of $E[Y_i(d)]$ under the weaker assumption that $\pi_{2d}(L_i, y)$ followed a parametric model such as logit $\pi_{2d}(L_i, y) = \alpha_0 + \alpha_1 L_i + \alpha_2 Y$.

3. CONCLUSION

The ATE parameter can be of greater public health interest than either the LATE or ITT parameter. We have proposed methods for setting bounds and for constructing estimators of the ATE parameter both in single active treatment trials and in bioequivalence trials. Both structural nested models and IPCW estimators can be applied to complex trials with randomized and nonrandomized time-dependent treatments, noncompliance, and dependent censoring with either failure time or repeated-measures outcomes (Robins 1989, 1993, 1994; Robins and Greenland 1994).

ADDITIONAL REFERENCES

Balke, A., and Pearl, J. (1993), "Bounds on Treatment Effects of Studies With Imperfect Compliance," Technical Report R-199-J, UCLA Cognitive Systems Laboratory.

Heyting, A., Tolboom, J. T. B. M., and Essers, J. G. A. (1992), "Statistical Handling of Drop-Outs in Longitudinal Clinical Trials," *Statistics in Medicine*, 11, 2043–2062.

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

——— (1994), "Learning About Social Programs From Experiments With Random Assignment of Treatment," Social Science Research Institute, 9505, University of Wisconsin.

Robins, J. M. (1987), "Addendum to A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect," *Computers and Mathematics with Applications*, 14, 923–945.

——— (1989), "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, D.C.: U.S. Public Health Service, pp. 113–159.

——— (1993), "Analytic Methods for Estimating HIV Treatment and Cofactor Effects," in *Methodological Issues of AIDS Mental Health Research*, eds. D. G. Ostrow and R. Kessler, New York: Plenum Publishing, pp. 213–290.

——— (1994), "Correcting for Noncompliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics*, 23, 2379–2412.

Robins, J. M., and Greenland, S. (1994), "Adjusting for Differential Rates of PCP Prophylaxis in High- Versus Low-Dose AZT Treatment Arms in an AIDS Randomized Trial," *Journal of the American Statistical Association*, 89, 737–749.

Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology—Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhäuser, pp. 297–331.

Robins, J. M., Rotnitzky, A., Zhao, L.-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Comment

James J. HECKMAN

Angrist, Imbens, and Rubin (AIR) apply the method of instrumental variables (IV) to estimate the local average treatment effect (LATE) of Imbens and Angrist (1994). Application of IV is routine, even for evaluation models with heterogeneous responses to treatment, so there is nothing novel or controversial about the method.

LATE is a controversial parameter because it is defined for an unobservable subpopulation. Its use as an evaluation parameter thus is of questionable value. More controversial yet is AIR's mischaracterization of the current state of knowledge about econometric models of simultaneity and selectivity. Econometrics has moved well beyond the (1943) model of Haavelmo and the simpler cases of the dummy endogenous variable model of Heckman (1978) to which the authors confine their attention in comparing their approach to econometric methods. It is interesting to contrast their commentary on econometric models for simultaneous equations with the commentary of scholars of causal analysis in science. For example, the distinguished philosopher Nancy Cartwright (1989) demonstrated how econometric methods for simultaneous equations elucidate causality, provide a coherent scheme for generating counterfactuals, and shed light on controversies in quantum mechanics.

This comment makes four points:

1. The "Rubin model" is a version of the widely used econometric switching regression model (Maddalla 1983; Quandt 1958, 1972, 1988). The Rubin model shares many features in common with the Roy model (Heckman and Honoré 1990; Roy 1951) and the model of competing risks (Cox 1962). It is a tribute to the value of the framework that it has been independently invented by different disciplines and subfields within statistics at different times.

2. Contrary to remarks by AIR, econometric work on simultaneous equations allows for variable responses to treatment, does not rely on arbitrary distributional assumptions, develops IV estimation methods for these models, examines the assumptions required to justify IV, and demonstrates that the assumptions required to use IV in the general case are very strong. This analysis is conducted within the context of clearly specified models of outcomes and regime selection that are motivated by behavioral theory. Econometricians make *weaker* mean independence assumptions rather than the strong independence assumptions made by AIR to identify their parameter.

3. The independence assumptions invoked by AIR are based on unspecified and implicit behavioral assumptions. These assumptions about behavior are very unattractive once they are clearly stated.

4. Econometric policy evaluation is designed to produce many counterfactuals from a common set of behavioral functions. Conditions required to nonparametrically identify this common set of functions are presented in the econometrics literature.

1. SWITCHING REGRESSION MODELS AND THE "RUBIN MODEL"

Counterfactuals are at the heart of any scientific study. Galileo was perhaps the first to use the thought experiment and the idealized method of controlled variation to define causal effects. Economists have used this method from the time of Alfred Marshall, who repeatedly used the principle of controlled variation in his *ceteris paribus* clauses. Since Haavelmo (1944), modern econometrics has been devoted to the construction and estimation of a broad array of counterfactuals. The construction of counterfactual states is the essence of econometric policy analysis (see Lucas and Sargent 1981).

Economists have used the following specific model of potential outcomes for at least 25 years. It has its origin in classical models of choice among discrete outcomes in mathematical psychology pioneered by Thurstone in the 1920s. (See Falmagne 1985 for an extensive bibliography; see also McFadden 1974 or Quandt 1972). There are two possible regimes, "0" or "1." (The generalization to an arbitrary number is trivial. I use two regimes to conform to the setup of AIR.) Associated with each regime is an outcome (Y^0 or Y^1). There is a rule that selects regimes. Let $D = 1$ if "1" is selected; $D = 0$ otherwise. Variables $\mathbf{X}$ determine outcomes in the following sense: $E(Y^1|\mathbf{X}) = \mu^1(\mathbf{X}), E(Y^0|\mathbf{X}) = \mu^0(\mathbf{X})$. Variables $\mathbf{X}$ and $\mathbf{Z}$ determine D in the following sense: $\Pr(D = 1|\mathbf{X}, \mathbf{Z})$ is a nontrivial function of both $\mathbf{X}$ and $\mathbf{Z}$. The "mysterious errors" that AIR censure in econometric work (see the last paragraph of their sec. 2) are (U^0, U^1) defined as $U^0 = Y^0 - E(Y^0|\mathbf{X}), U^1 = Y^1 - E(Y^1|\mathbf{X})$. (There are more general models, as considered in Heckman and Robb 1985, but for the sake of brevity I consider only the simplest case.) These "errors" are well defined as long as $E(Y^0) < \infty, E(Y^1) < \infty$. If (U^0, U^1) are independent across observations "SUTVA" holds.

Only one regime is observed at any time. Let Y be the observed outcome. $\Delta = Y^1 - Y^0$. Then

$$Y = DY^1 + (1 - D)Y^0 = Y^0 + D(Y^1 - Y^0) = Y^0 + D\Delta, \tag{1}$$

James J. Heckman is the Henry Schultz Distinguished Service Professor, Department of Economics, University of Chicago. This research was supported by National Science Foundation Grant SBR-93-048-0211 and a grant from the Russell Sage Foundation.

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

so

$$Y = \mu^0(\mathbf{X}) + D(\mu^1(\mathbf{X}) - \mu^0(\mathbf{X}) + U^1 - U^0) + U^0. \quad (2)$$

The "gain" from going from "0" to "1" is $\Delta = Y^1 - Y^0$. This is what AIR and previous authors in econometrics call the "causal effect" of D on Y. If superscript "1" refers to demand and "0" to supply, we obtain the disequilibrium markets model of Quandt (1972, 1988) where

$$D = 1(Y^1 \leq Y^0); = 0 \quad \text{otherwise}.$$

"1" is the logical indicator variable ($1(a) = 1$ if a holds). Letting Y^1 be the market wage and Y^0 the value of nonmarket time, we obtain the model of the value of wages and of nonmarket time presented by Gronau (1974) and Heckman (1974), where $D = 1(Y^1 \geq Y^0)$. The same model appears in studies of search unemployment (Flinn and Heckman 1982). If Y^1 and Y^0 are times of death in a competing risk setup, then $D = 1(Y^1 \leq Y^0)$. Amemiya (1985), Heckman and Robb (1985), and Willis and Rosen (1979) considered more general models for D. Applications of this basic model of potential outcomes are legion in economics. Analyses are available both for the case where D is observed and where it is not. (See the numerous references in Amemiya 1985.)

2. MEAN EFFECT OF TREATMENT ON THE TREATED: A "CAUSAL" PARAMETER

A basic program evaluation parameter widely used in economics is the mean effect of treatment on the treated:

$$E(Y^1 - Y^0 | D = 1, \mathbf{X}) = E(\Delta | D = 1, \mathbf{X}). \quad (3)$$

This parameter tells us what, on average, persons of characteristics $\mathbf{X}$ who actually participate in regime 1 gain from a switch from regime "0" to "1." Unlike LATE, this parameter is defined for an observable subpopulation. It is a parameter that corresponds more closely to the coefficient on the dummy variable in nonlinear simultaneous equation (2) with variable treatment effect than does LATE. Using means of nonparticipants to estimate $E(Y^0|D = 1, X)$ gives rise to selection bias if $E(Y^0|D = 1, \mathbf{X}) - E(Y^0|D = 0, \mathbf{X}) \neq 0$. Equivalently, $E(U^0|\mathbf{X}, D = 1) \neq E(U^0|\mathbf{X}, D = 0)$. I discuss identification and estimation of this parameter because there is a vast literature on it in economics, and the lessons from an analysis of this parameter apply directly to LATE.

Equation (2) is a *variable effect* or random coefficient model. The response to treatment (the term multiplying D) varies among persons with identical $\mathbf{X}$, unless $U^1 = U^0$. That special case is the dummy endogenous variable model of Heckman (1978) discussed by AIR. We may reparameterize Equation (2) in terms of $E(\Delta|D = 1, \mathbf{X})$. The reparameterization writes

$$Y = \mu^0(\mathbf{X}) + D(E(\Delta|\mathbf{X}, D = 1)$$
$$+ \{U^0 + D(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1))\}.$$

The term in braces is an interpretable "error term." If D were orthogonal to this term, then simple means for each $\mathbf{X}$ group would identify the parameter of interest. (For each $\mathbf{X}$, subtract the mean for $D = 0$ from the mean for $D = 1$). Selection bias makes D nonorthogonal to U^0. D is orthogonal to the second error component in the braces ($E(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1)|\mathbf{X}, D = 1) = 0$)), but not to the first component.

The method of IV has been applied to identify this parameter under general conditions (see Heckman and Robb 1985, 1986). Suppose that Z is distinct from X (i.e., does not appear directly in (2)) and satisfies the following *mean independence* conditions:

$$E(U^0|\mathbf{X}, \mathbf{Z}) = 0 \quad \text{(A-1)}$$

$$E(U^1 - U^0 - E(U^1 - U^0|\mathbf{X}, D = 1)|\mathbf{X}, \mathbf{Z}, D = 1) = 0. \quad \text{(A-2)}$$

An alternative and equivalent way to write condition (A-2) is:

$$E(\Delta|\mathbf{X}, \mathbf{Z}, D = 1) = E(\Delta|\mathbf{X}, D = 1). \quad \text{(A-2')}$$

Finally, restate the condition that both $\mathbf{Z}$ and $\mathbf{X}$ determine D as an assumption:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}) \neq \Pr(D = 1|\mathbf{X}) \quad \text{and} \quad \Pr(D = 1|\mathbf{X}, \mathbf{Z}) \quad \text{(A-3)}$$

is a nontrivial function of $\mathbf{Z}$. Then

$$E(Y|\mathbf{X},\mathbf{Z}) = \mu^0(\mathbf{X}) + E(\Delta|\mathbf{X}, D = 1)\Pr(D = 1|\mathbf{X},\mathbf{Z}). \quad (4)$$

If for each $\mathbf{X}$ there are at least two distinct values of $\mathbf{Z}$ such that $\Pr(D = 1|\mathbf{X}, \mathbf{Z}') \neq \Pr(D = 1|\mathbf{X}, \mathbf{Z}'')$, then we may evaluate (4) at all values of X to obtain

$$E(\Delta|\mathbf{X}, D = 1)$$
$$= \left(\frac{E(Y|\mathbf{X}, \mathbf{Z}') - E(Y|\mathbf{X}, \mathbf{Z}'')}{\Pr(D = 1, \mathbf{X}, \mathbf{Z}') - \Pr(D = 1|\mathbf{X}, \mathbf{Z}'')} \right). \quad (5)$$

If for some $\mathbf{X}$ values there are not distinct values for $\Pr(D = 1|\mathbf{X}, \mathbf{Z})$ for two or more values of Z, then the parameter is not identified at those values. Replacing population objects with sample mean analogs produces the IV estimator.

Observe that only mean independence is required—not full independence, as assumed by AIR. AIR are able to test their identifying assumptions because they invoke much stronger conditions than are required to identify their parameter. Minimal identifying assumptions cannot be tested. (Heckman and Robb 1985). Note further that no arbitrary and untestable monotonicity condition is needed—just a condition that guarantees that the denominator of (5) is not zero for the particular value of $\mathbf{X}$. Parenthetically, monotonicity is not required in classical discrete choice theory either. Also, even in the original dummy endogenous variable analysis it is recognized that the second assumption of AIR's Equation (4) is not needed to apply IV.

3. ARE THE ASSUMPTIONS VALID?

A central focus in modern econometrics is the development of explicit behavioral models relating the "errors" and choices made by agents. This is critical to developing and justifying any econometric evaluation strategy. Therefore, it is surprising to read in AIR that econometricians do not clearly state assumptions like (A-1)–(A-3) or worry about the justification for them. Hansen and Sargent (1991), and Heckman and Robb (1985, 1986) are just some of the authors who have built explicit behavioral models to justify exogeneity and noncausality assumptions.

Assumption (A-1) is conventional; Assumption (A-2) (or A-2′) is not. (A-2) is satisfied if $U^1 - U^0 = 0, (Y^1 - Y^0 = \mu^1(\mathbf{X}) - \mu^0(\mathbf{X}))$, so that conditional on **X** there is no treatment response heterogeneity. It is also satisfied in the case of heterogeneous response to treatment when the rule governing participation in a regime conditional on **Z**, and **X** does not depend on $Y^1 - Y^0$:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}, Y^1 - Y^0) = \Pr(D = 1|\mathbf{X}, \mathbf{Z}), \quad \text{(6a)}$$

provided that $Y^1 - Y^0$ conditional on X is not perfectly forecastable by **Z**. This is a Granger noncausality condition routinely used in econometrics and explicitly presented in this context by Heckman and Robb (1985, 1986). Alternatively, in terms of the "mysterious" unobservables to which AIR object, the condition is:

$$\Pr(D = 1|\mathbf{X}, \mathbf{Z}, U^1 - U^0) = \Pr(D = 1|\mathbf{X}, \mathbf{Z}), \quad \text{(6b)}$$

provided that $U^1 - U^0$ conditional on **X** is not perfectly forecastable by **Z**. AIR call this non-causality condition "ignorability." If $U^1 - U^0$ is perfectly forecastable by **Z**, conditional on **X**, then (A-2) would be violated.

In general, the extra conditioning on **Z** causes (A-2) to be violated although it is trivially satisfied only if conditioning is done on X and D. The behavioral assumption justifying (6a) and (6b) requires that the relevant decision makers do not make decisions about which regime is selected using information on the outcomes of the regime that cannot be forecast by X and Z. In most situations, persons making decisions have more information about the outcomes than the statisticians studying them. This makes assumption (6a) or (6b) questionable in such cases.

This assumption is definitely not satisfied in the competing risks model, in the Gronau–Heckman market wage–nonmarket wage model, in the Roy model (Heckman and Honoré 1990), or in most versions of the switching regressions model. These limitations on the application of the IV method were spelled out by Heckman and Robb (1985, 1986) and later reiterated by Heckman (1995). In the switching regression context, they were discussed by Quandt (1988). Although space limitations preclude the full development of the point, IV estimation of LATE requires the same stringent behavioral assumptions.

The draft lottery number cited by AIR as a valid instrument is unlikely to satisfy (A-2) or (A-3) and thus is not likely to be a valid instrument. Consider the application of IV by Angrist (1990) that AIR discuss. The potential outcomes are earnings if persons serve in the military or if they do not. Persons who get a high number are virtually guaranteed that they are exempt from service. Those persons with a high number who nonetheless volunteer to go to the Army perceive a high gain from doing so. If those perceptions are related to the potential outcomes and are based on private information that cannot be fully predicted by **X** and **Z**, then the lottery number is not a proper instrument. In addition, persons with high numbers are likely to receive more job training, because their likelihood of being drafted is reduced and firms have less likelihood of losing them. Then **Z** is an **X**, and the exclusion assumption is violated. Because of the stringent nature of the required assumptions, most economists have been very cautious about using IV to identify the parameters of switching models. Sometimes, however, application of IV can be justified in the context of heterogeneous treatments. Robinson (1989), using a test proposed by Heckman and Robb (1985, 1986), demonstrated that IV methods produce appropriate estimates for estimating the "causal effect" of unions on wages; that is, the union–non-union wage differential. Robinson's evidence is surprising because it indicates that union membership is not based on unobserved components of union wage differentials not predicted by the crude **X** and **Z** available to him.

A major difference between the approach taken by AIR and that used by econometricians is that the latter go to much greater depth in justifying the behavioral assumptions that are implicit in the statistical assumptions. It is disappointing to see an entire literature in econometrics that develops explicit models designed to test and justify (A-1)–(A-3), or other identifying assumptions, ignored by AIR in their discussion of the econometrics literature.

4. IDENTIFICATION UNDER MORE GENERAL CONDITIONS FOR A VARIETY OF PARAMETERS

Heckman and Honoré (1989, 1990) presented conditions for identifiability of the full distributions of outcomes in the competing risks and Roy models. Heckman (1990) considered nonparametric identifiability in more general models. Björklund and Moffitt (1986) considered estimation of the more general models under specific distributional assumptions. Those authors demonstrate that other methods besides IV estimate behaviorally interesting parameters under more behaviorally plausible conditions. These more general models produce identification of a large array of distinct counterfactuals—a central goal of structural econometric policy evaluation—and do not focus on just one special parameter.

ADDITIONAL REFERENCES

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Bjorklund, A., and Moffitt, R. (1987), "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69, 42–49.

Falmagne, J. (1985), *Elements of Psychophysical Theory*, Oxford, U.K.: Clarendon Press.

Flinn, C., and Heckman, J. (1982), "New Methods for Analyzing Structural

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 affect 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):

$$\beta_1^{\mathrm{IV}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}, \tag{1}$$

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 β_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:

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

where k is the fraction of the total variance of D that arises from the "between" and β_1^{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, β_1^{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 Nonexperimental 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: Spring-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 illumined 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 observed 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

Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

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

Table 1. An Experiment Encouraging Exercise for Lung Disease

Subject	Exercise encouraged Z_i	Exercise performed $D_i(Z_i)$	FEV response without exercise $Y_i(0)$	Observed FEV response $\tilde{Y}_i$
1	1	1	5	8
2	1	1	4	7
3	1	1	3	6
4	1	0	2	2
5	1	0	1	1
6	0	1	5	8
7	0	0	4	4
8	0	0	3	3
9	0	0	2	2
10	0	0	1	1

the 1.2 determined previously by comparing the encouraged and control groups, but it attributes the entire 1.2 to the increase in exercise in the encouraged group; that is, it divides 1.2 by the mean difference in exercise, $(1+1+1+0+0)/5-(1+0+0+0+0)/5=2/5$, so the estimate is $1.2/(2/5)=3$.

The point made later is that this argument has nothing to do with means and quickly extends to sturdier estimates, such as the Hodges–Lehmann (HL) (Hodges and Lehmann 1963). Also, with an instrumental variable, exact permutation inferences about an additive effect are obtained using the random assignment of encouragement. Moreover, sensitivity analysis is straightforward in nonrandomized or observational studies. (See Hollander and Wolfe 1973 or Lehmann 1975 for discussion of the standard forms of the HL estimate, and see Maritz 1995, secs. 1 and 8.1 for discussion of the broad scope of HL estimates with various technical results.)

3. THE HODGES–LEHMANN ESTIMATE USING AN INSTRUMENTAL VARIABLE IN A RANDOMIZED EXPERIMENT

Following AIR, assume the stable unit treatment value assumption (SUTVA), the exclusion restriction, and the nonzero average effect of Z on D. For subject i, write $\tilde{D}_i$ for the observed level of exercise, $\tilde{D}_i = Z_i \cdot D_i(1) + (1-Z_i)D_i(0)$, and write $\tilde{Y}_i$ for the observed outcome, $\tilde{Y}_i = \tilde{D}_i \cdot Y_i(1) + (1-\tilde{D}_i)Y_i(0)$. The model of an additive effect asserts that $Y_i(1) - Y_i(0) = \tau$ for all i, so $\tilde{Y}_i = Y_i(0) + \tau\tilde{D}_i$, as in Table 1 where $\tau = 3$. In Section 5, it will be seen that the additive model need not hold for all i, that it suffices that additivity holds for subjects who change treatments in response to encouragement, but it is easier to discuss this separately. Write $\mathbf{Z}, \tilde{\mathbf{D}}, \tilde{\mathbf{Y}}$, and $\mathbf{Y}_0$ for the N-dimensional vectors of Z_i's, $\tilde{D}_i$'s, $\tilde{Y}_i$'s, and $Y_i(0)$'s. Write M for the number of encouraged subjects, $M = \mathbf{Z}^T\mathbf{Z}$. In Table 1, $M = 5$. Write B for the set containing the possible treatment assignments, so B contains $\binom{N}{M}$ vectors of dimension N with M coordinates equal to 1 and $N-M$ coordinates equal to zero. In a randomized experiment, $\mathbf{Z}$ is picked from B at random; that is, $\text{prob}(\mathbf{Z}=\mathbf{z}) = \binom{N}{M}^{-1}$ for each $\mathbf{z} \in B$.

Let $t(\mathbf{Z}, \tilde{\mathbf{Y}})$ be a statistic used to compare the encouraged and control groups. For instance, $t(\mathbf{Z}, \tilde{\mathbf{Y}})$ might be the difference in sample means, say $t_M(\mathbf{Z}, \tilde{\mathbf{Y}}) = \mathbf{Z}^T\tilde{\mathbf{Y}}/M - (1-\mathbf{Z})^T\tilde{\mathbf{Y}}/(N-M)$, or Wilcoxon's rank sum statistic, $t_{\mathrm{W}}(\mathbf{Z}, \tilde{\mathbf{Y}}) = \mathbf{Z}^T\text{rank}(\tilde{\mathbf{Y}})$, where $\text{rank}(\tilde{\mathbf{Y}})$ is the N-dimensional vector of ranks of the $\tilde{\mathbf{Y}}$ with average ranks used for ties, or the difference between the trimean in the encouraged group and the trimean in the control group, say $t_{\mathrm{T}}(\mathbf{Z}, \tilde{\mathbf{Y}})$. (Recall that the trimean is the sum of the upper and lower quartiles plus twice the median divided by four.)

Let $\bar{\bar{t}}$ be the expectation of $t(\mathbf{Z}, \mathbf{Y}_0)$ over the randomization distribution of $\mathbf{Z}$; that is, the average of $t(\mathbf{z}, \mathbf{Y}_0)$ over the $\binom{N}{M}$ choices $\mathbf{z} \in B$. For the difference in means, $t_{\mathrm{M}}(\mathbf{Z}, \mathbf{Y}_0)$ has expectation $\bar{\bar{t}}_{\mathrm{M}} = 0$. For the rank sum, $t_{\mathrm{W}}(\mathbf{Z}, \mathbf{Y}_0)$ has expectation $\bar{\bar{t}}_{\mathrm{W}} = M(N+1)/2$. For the difference of trimeans, $t_{\mathrm{T}}(\mathbf{Z}, \tilde{\mathbf{Y}})$ has expectation $\bar{\bar{t}}_{\mathrm{T}} = 0$ if $M = N/2$ and $\bar{\bar{t}}_{\mathrm{T}} \to 0$ as $M, N-M \to \infty$ whether or not $M = N/2$.

Now, $\mathbf{Y}_0 = \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau$. The HL estimate using an IV is defined to be the value $\hat{\tau}$ such that $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau})$ is as close as possible to $\bar{\bar{t}}$. For some estimators such as the difference in means, $\hat{\tau}$ may be determined by solving $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) = \bar{\bar{t}}$. For estimates based on rank statistics that move in discrete steps, there may be no $\hat{\tau}$ such that $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) = \bar{\bar{t}}$ exactly; then, following HL, $\hat{\tau}$ is defined as the average of the smallest value that is too large and the largest value that is too small, namely

$$\hat{\tau} = \frac{\sup\{\tau\colon t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) > \bar{\bar{t}}\} + \inf\{\tau\colon t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau) < \bar{\bar{t}}\}}{2}.$$

Here $\hat{\tau} = \infty$ if there is no finite τ such that $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau) < \bar{\bar{t}}$ and $\hat{\tau} = -\infty$ if there is no finite $\hat{\tau}$ such that $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau) > \bar{\bar{t}}$.

Write $\hat{\tau}_{\mathrm{M}}, \hat{\tau}_{\mathrm{W}}$, and $\hat{\tau}_{\mathrm{T}}$ for the instrumental HL estimates based on $t_{\mathrm{M}}, t_{\mathrm{W}}$, and t_{T}. For the difference in means, simple algebra shows $t_{\mathrm{M}}(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) = \bar{\bar{t}} = 0$ if and only if $\hat{\tau}_{\mathrm{M}}$ is the usual instrumental variable estimator discussed by AIR. If encouragement always determines the treatment so $\tilde{\mathbf{D}} = \mathbf{Z}$, then $\hat{\tau}_{\mathrm{M}}$ is the encouraged minus-control difference in sample means, $\hat{\tau}_{\mathrm{W}}$ is the usual HL estimate associated with the rank sum statistic, and $\hat{\tau}_{\mathrm{T}}$ is the difference in trimeans. In short, the estimate $\hat{\tau}$ generalizes both the usual IV estimate and the usual HL estimate.

In the example in Table 1, subtracting 3 from each subject who exercised sets all three statistics, $t_{\mathrm{M}}, t_{\mathrm{W}}$, and t_{T}, equal to their null expectations, so $\hat{\tau}_{\mathrm{M}} = \hat{\tau}_{\mathrm{W}} = \hat{\tau}_{\mathrm{T}} = 3$ in this particular case. This is exceptional and reflects the perfect balance of the $Y_i(0)$'s in this constructed example. If $\tilde{Y}_1$ in Table 1 were replaced by an extremely large positive value, then $\hat{\tau}_{\mathrm{M}}$ would increase dramatically, $\hat{\tau}_{\mathrm{W}}$ would increase slightly, and $\hat{\tau}_{\mathrm{T}}$ would continue to equal 3.

Exact inference about τ may be based on the randomization distribution of $t(\mathbf{Z}, \mathbf{Y}_0)$ where $\mathbf{Y}_0$ is fixed. Consider first testing the null hypothesis that H_0: $\tau = \tau^*$. Under the null hypothesis, the fixed responses in the absence of exercise, $\mathbf{Y}_0$, are equal to $\tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau^*$, which may be computed from the data. The one-sided randomization significance level is the proportion of treatment assignments $\mathbf{z} \in B$ giving a larger value of the test statistic than observed, $|\{\mathbf{z} \in B\colon t(\mathbf{z}, \mathbf{Y}_0) \geq t(\mathbf{Z}, \mathbf{Y}_0)\}|/\binom{N}{M}$, where $|A|$ denotes the number of elements of the set A. A one-sided confidence interval for τ is obtained by determining all values of τ not rejected by such a test. A two-sided 95% confidence interval is the intersection of two one-sided 97.5% intervals. It is notable that the test of H_0: $\tau = 0$ is identical to the usual randomization test of no effect, as discussed by Fisher (1935) or Kempthorne (1952), but this is not true for $\tau^* \neq 0$.

For instance, in Table 2, to test the false hypothesis $H_0\colon \tau = 1.5$ using the rank sum test with an instrumental variable, one computes $\tilde{\mathbf{Y}} - 1.5\tilde{\mathbf{D}} = (6.5, 5.5, 4.5, 2, 1, 6.5, 4, 3, 2, 1)$ with ranks $(9.5, 8, 7, 3.5, 1.5, 9.5, 6, 5, 3.5, 1.5)$, so the rank sum is $9.5 + 8 + 7 + 3.5 + 1.5 = 29.5$. Allowing for the ties, the null expectation and variance of the rank sum are 27.5 and 22.5 yielding a standardized deviate of $(29.5 - 27.5)/\sqrt{22.5} = .42$, so the hypothesis H_0: $\tau = 1.5$ is not rejected. Without ties, the familiar exact distribution of the rank sum statistic may be used.

In short, permutation tests, confidence intervals, and HL estimates all use the null distribution of $t(\mathbf{Z}, \mathbf{Y}_0)$, which is the usual randomization distribution (Fisher 1935; Kempthorne 1952, sec. 8.2). If encouragement itself $\mathbf{Z}$ had an additive effect, then one would have $\mathbf{Y}_0 = \tilde{\mathbf{Y}} - \mathbf{Z}\tau$, and the usual procedures for permutation inference would result. What is new with the IV is that exercise $\tilde{\mathbf{D}}$ and not encouragement $\mathbf{Z}$ has the additive effect, so the permutation inference is based on $\mathbf{Y}_0 = \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau$, but otherwise permutation methods are unchanged. As it turns out, these considerations extend immediately for sensitivity analysis in observational studies.

4. SENSITIVITY ANALYSIS USING INSTRUMENTAL VARIABLES IN OBSERVATIONAL STUDIES

In the experiment in Table 1, random assignment of encouragement tended to balance the distribution of $Y_i(0)$ in encouraged ($Z_i = 1$) and control ($Z_i = 0$) groups. In an observational study or nonrandomized experiment, subjects might have differing chances of receiving encouragement to exercise; that is, it may be quite wrong to assume that $\mathrm{prob}(\mathbf{Z} = \mathbf{z}) = \binom{N}{M}^{-1}$ for each $\mathbf{z} \in B$. Perhaps the severely ill would be less likely to receive encouragement than the less severely ill.

It is possible to study the sensitivity of permutation inferences to departures from random assignment of treatments (see, for instance, Rosenbaum 1993, 1995). These techniques replace $\mathrm{prob}(\mathbf{Z} = \mathbf{z}) = \binom{N}{M}^{-1}$ with a range of distributions of treatment assignments, thereby obtaining a range of null distributions for $t(\mathbf{Z}, \mathbf{Y}_0)$. Using these techniques with an IV is straightforward; one calculates $\mathbf{Y}_0 = \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\tau$ as in Section 3 and applies the sensitivity analysis to the result. For instance, the sensitivity analysis gives not one null expectation $\bar{\bar{t}}_{\mathrm{W}} = M(N+1)/2$ for the rank sum statistic, but rather a range of expectations $[\bar{\bar{t}}_{\mathrm{W,low}}, \bar{\bar{t}}_{\mathrm{W,high}}]$ depending on a sensitivity parameter Γ. This yields a range of instrumental HL estimates obtained by approximately solving $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) = \bar{\bar{t}}_{\mathrm{W,low}}$ and $t(\mathbf{Z}, \tilde{\mathbf{Y}} - \tilde{\mathbf{D}}\hat{\tau}) = \bar{\bar{t}}_{\mathrm{W,high}}$.

5. AVOIDING SPECULATION ABOUT SUBJECTS WHO IGNORE ENCOURAGEMENT

AIR carefully focus attention on subjects who do not ignore encouragement; see, for instance, their proposition 1. They call subjects who ignore encouragement "always-takers" if $D_i(1) = D_i(0) = 1$ or "never-takers" if $D_i(1) = D_i(0) = 0$. They argue, in effect, that one can say little about subjects who ignore encouragement because nothing that the experimenter does will change the treatment they receive. This final section briefly observes that the argument in Section 3 continues to hold if nothing is assumed about subjects who ignore encouragement.

Following AIR, consider a randomized experiment and assume that Z is an instrumental variable in the sense of their definition 3, so $D_i(1) \geq D_i(0)$. In addition, assume that the treatment has an additive effect for compliers only; that is, $Y_i(1) - Y_i(0) = \tau$ whenever $D_i(1) = 1 > 0 = D_i(0)$. No assumption is made about $Y_i(1) - Y_i(0)$, when $D_i(1) = D_i(0)$. Let τ^* be a hypothesized value for τ. Then the adjusted responses are

$$\tilde{Y}_i - \tilde{D}_i\tau^* = \begin{cases} Y_i(1) - \tau^* \quad \text{if } D_i(1) = D_i(0) = 1, \\ Y_i(0) + (\tau - \tau^*)Z_i \\ \quad \text{if } D_i(1) = 1 > 0 = D_i(0), \\ Y_i(0) \quad \text{if } D_i(1) = D_i(0) = 0. \end{cases}$$

Note first that the adjusted responses $\tilde{Y}_i - \tilde{D}_i\tau^*$ will be independent of encouragement Z_i if and only if the hypothesized τ^* equals the true τ. Also, if $\tau^* < \tau$, then the adjusted responses $\tilde{Y}_i - \tilde{D}_i\tau^*$ for encouraged subjects ($Z_i = 1$) will tend to be somewhat higher than those for control ($Z_i = 0$) subjects, and conversely if $\tau^* > \tau$. As a consequence, to render the adjusted responses independent of encouragement, one must have the correct τ^*, and a test statistic such as the rank sum statistic that is consistent when one distribution is stochastically larger than another will, in sufficiently large sample sizes, reject any fixed $\tau^* \neq \tau$, thereby yielding consistent tests, confidence intervals, and point estimates. In short, the procedures in Section 3 describe subjects who comply with no assumptions about those who ignore encouragement.

ADDITIONAL REFERENCES

Hodges, J., and Lehmann, E. (1963), "Estimates of Location Based on Rank Tests," *Annals of Mathematical Statistics*, 34, 598–611.

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

© 1996 American Statistical Association
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 related 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.