# Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity

Joshua D. ANGRIST and Guido W. IMBENS*

Two-stage least squares (TSLS) is widely used in econometrics to estimate parameters in systems of linear simultaneous equations and to solve problems of omitted-variables bias in single-equation estimation. We show here that TSLS can also be used to estimate the average causal effect of variable treatments such as drug dosage, hours of exam preparation, cigarette smoking, and years of schooling. The average causal effect in which we are interested is a conditional expectation of the difference between the outcomes of the treated and what these outcomes would have been in the absence of treatment. Given mild regularity assumptions, the probability limit of TSLS is a weighted average of per-unit average causal effects along the length of an appropriately defined causal response function. The weighting function is illustrated in an empirical example based on the relationship between schooling and earnings.

KEY WORDS: Instrumental variables; Rubin causal model; Schooling; Wages.

## 1. INTRODUCTION

Econometricians and statisticians have developed a variety of techniques for the estimation of systems of linear simultaneous equations. The simplest and most commonly used of these techniques is the class of instrumental variables (IV) estimators, of which two-stage least squares (TSLS) is the most important special case. IV and TSLS were developed in early research on simultaneous equations estimation (by Wright [1928] and Theil [1958], among others), and both estimators are now described in every econometrics textbook (e.g., Theil 1971). These techniques are typically introduced as solutions to the problem of "simultaneous equations bias." More generally, however, IV provides a powerful and flexible estimation strategy that can be used to tackle the problem of omitted-variables bias in a wide range of single-equation regression applications, such as models with mismeasured regressors (Durbin 1954) and the estimation of treatment effects in manpower training programs (Heckman and Hotz 1989; Heckman and Robb 1985).

The use of IV and TSLS to estimate treatment effects is not limited to econometrics. Permutt and Hebel (1989) used TSLS to estimate the effect of maternal smoking on birth weight. Hearst, Newman, and Hulley's (1986) use of the draft lottery to estimate the effect of Vietnam-era military service on civilian mortality is another epidemiological application of IV. Angrist and Krueger (1992) used IV to estimate the effect of children's age at school entry on their ultimate educational attainment. The Powers and Swinton (1984) "encouragement design" used to study the effect of test preparation on graduate record examination (GRE) scores, discussed by Holland (1988), also leads naturally to IV and TSLS estimators. Finally, the analysis used by Robins (1989, sec. 18) and Robins and Tsiatis (1991) to correct for noncompliance in clinical trials is an application of instrumental variables to experimental data.

Evaluation research in econometrics and applications of IV and TSLS in other fields typically rely on regression models with constant coefficients. The regression approach to evaluation postulates a hypothetical linear response function. In contrast, the statistical literature on evaluation has been strongly influenced by Rubin's (1974, 1977) model for causal inference using counterfactual outcomes. (The ideas behind this framework date back to Neyman and Fisher; see Rubin 1990 for historical details.) In a methodological paper related to this one (Imbens and Angrist 1994), we discuss the use of IV to estimate average causal effects in Rubin's causal model for binary treatments, as well as how to estimate the sampling variance of IV estimates of treatment effects. In another recent paper (Angrist, Imbens, and Rubin 1995), we present a detailed analysis of the conceptual issues and problems that arise when IV is used to estimate the average causal effect of a binary treatment in the Rubin causal model, together with a survey of evaluation research in econometrics and statistics.

In this article, we show how TSLS can be used to estimate average causal effects in a version of Rubin's causal model that allows for variable treatment intensity, multiple instruments, and covariates. In particular, we show that TSLS applied to a causal model with variable treatment intensity and nonignorable treatment assignment identifies a weighted average of per-unit treatment effects along the length of a causal response function. Our results do not hinge on linearity of the relationships between response variables, treatment intensities, and instruments.

A second contribution of this article is to illustrate the causal response weighting function in an empirical example based on the work by Angrist and Krueger (1991), using IV and TSLS to estimate the effect of years of schooling on earnings. The weighting function in the application is of in-

* Joshua D. Angrist is Senior Lecturer, Department of Economics, Hebrew University, Jerusalem 91095, Israel. Guido W. Imbens is Associate Professor, Department of Economics, Harvard University, Cambridge, MA 02138.
The authors thank Don Rubin for raising the question that this article answers, and seminar participants at the Harvard-MIT Econometrics Workshop and the University of Wisconsin for helpful comments. Four referees and an associate editor also made helpful comments. Financial support from the National Science Foundation (SES-9122627) and the Nederlandse Organisatie Voor Wetenschappelijk Onderzoek is also gratefully acknowledged. Part of the article was written while the second author was visiting the Hebrew University Department of Economics.


Journal of the American Statistical Association
June 1995, Vol. 90, No. 430, Applications and Case Studies

terest because the relationship between schooling and earnings is one of the most important empirical regularities in economics. More generally, the weighting formula can help researchers understand which observations are contributing to a particular estimate, and the formula provides a causal interpretation for some of the simple estimators commonly used in applied research.

## 2. APPLICATION: THE EFFECT OF COMPULSORY SCHOOL ATTENDANCE ON EARNINGS

The theory of human capital (Becker 1964) says that years of schooling can be treated much the same way as an investment in physical capital, yielding a rate of return something like an interest rate. The empirical counterpart of this theory is the "human capital earnings function," a semilogarithmic regression of earnings on schooling and other covariates. An example is the following model for microdata:

$$Y = \gamma_0 + X_1\gamma_1 + \rho S + \varepsilon, \qquad (1)$$

where $Y$ is the log of weekly earnings, $\gamma_0$ is a constant, $X_1$ is a row vector of covariates, $\gamma_1$ a vector of coefficients, $S$ is years of schooling, and the coefficient $\rho$ is the approximate percentage return to a year of schooling. Equation (1) is sometimes augmented with an equation describing how schooling is related to the covariates, $X_1$, and additional covariates, $X_2$:

$$S = \delta_0 + X_1\delta_1 + X_2\delta_2 + \eta. \qquad (2)$$

Econometricians typically assume that a semilogarithmic response function for earnings is a reasonably good approximation to the true earnings function. Ordinary least squares (OLS) applied to (1), however, may lead to biased estimates of $\rho$ even if the true response function is linear. The reason is that schooling is determined by individual choices under constraints. For example, the literature on schooling and earnings has devoted considerable attention to the problem of "ability bias" in estimates of the economic return to schooling. This is a form of omitted-variables bias that would arise if more able individuals in the labor market get more schooling, perhaps because of better access to capital markets. The observed positive correlation between schooling and earnings would then partly reflect the fact that those with more schooling have higher earnings potential. In terms of Equations (1) and (2), ability is common to the error terms, $\varepsilon$ and $\eta$, and so these error terms are correlated.

One infeasible solution to this problem is to conduct an experiment in which schooling is randomly assigned. Random assignment would eliminate the correlation between schooling and ability or unobserved earnings potential. Even in the absence of a true experiment, a "natural experiment" may generate instrumental variables that effectively do the same thing.

Instrumental variables are variables related to the outcome of interest solely through the treatment of interest. For example, in two recent papers, Angrist and Krueger (1991, 1992) showed that students' quarter of birth interacts with compulsory attendance laws and age at school entry to generate variation in years of completed schooling. State compulsory attendance laws typically require students to enter school in the fall of the year in which they turn 6, but allow students to drop out of school when they reach age 16. This induces a relationship between quarter of birth and educational attainment, because students born in the first quarter of the year enter school at an older age than do students born in later quarters. Students who enter school at an older age thus are allowed to drop out after having completed less schooling than students who enter school at a younger age. If students' quarter of birth is correlated with earnings solely because it is correlated with schooling, then it is an instrument for schooling in an earnings equation.

One way to convert this idea into an estimation strategy is to compare the education and earnings of people born in the first quarter to the education and earnings of people born in a later quarter, say the fourth. This leads to the simplest possible IV estimator: Wald's method of fitting straight lines (Durbin 1954). As an example, calculations underlying Wald estimates based on a first quarter/fourth quarter comparison for men are laid out in Table 1. Panel A of the table shows results tabulated from data on the wages and earnings of men in the 1970 Census, and Panel B shows results tabulated using data from the 1980 Census. In both data sets, men born in the first quarter earn slightly less and have slightly less schooling than men born in the fourth quarter. The ratio of differences in earnings to differences in schooling generates a Wald estimate of the return to schooling of 5.3% using the 1970 Census and 8.9% using the 1980 Census.

Durbin (1954) showed that under the null hypothesis that the OLS estimate is consistent, the sampling variance of the

*Table 1. Compulsory School Attendance*

| | (1) Born in 1st quarter of year | (2) Born in 4th quarter of year | (3) Difference (std. error) (1) − (2) |
|---|---|---|---|
| *Panel A: Wald Estimates for 1970 Census—Men Born 1920–1929*[a] | | | |
| ln (weekly wage) | 5.1485 | 5.1578 | −.00935 (.00374) |
| Education | 11.3996 | 11.5754 | −.1758 (.0192) |
| Wald est. of return to education | | | .0531 (.0196) |
| OLS est. of return to education[b] | | | .0797 (.0005) |
| *Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939* | | | |
| ln (weekly wage) | 5.8916 | 5.9051 | −.01349 (.00337) |
| Education | 12.6881 | 12.8394 | −.1514 (.0162) |
| Wald est. of return to education | | | .0891 (.0210) |
| OLS est. of return to education | | | .0703 (.0005) |

[a] The sample size is 122,223 in Panel A and 162,515 in Panel B. Each sample consists of men born in the first and fourth quarters of the year in the United States who had positive earnings in the year preceding the survey. The 1980 Census sample is drawn from the 5% sample, and the 1970 Census sample is from the state, county, and neighborhoods 1% samples. A detailed description of the data sets is provided in the Appendix to Angrist and Krueger (1991).

[b] The OLS return to education was estimated from a bivariate regression of log weekly earnings on years of education in a sample of men born in the first and fourth quarters.

difference between a Wald and OLS estimates is the difference in their variances. Ignoring the small sampling error of the OLS estimates, the Wald estimates are within sampling error of the OLS estimates (8% and 7% in the two Census data sets). Thus the instrumental variables estimates seem to support a conclusion that the OLS estimates are not biased by unobserved ability in the error term. (A detailed case for this claim is made in Angrist and Krueger 1991, which also discusses strategies to control for age effects when using quarter of birth as an instrument for schooling.)

The general formula for an IV estimate is $(\mathbf{Z}'\mathbf{W})^{-1}\mathbf{Z}'\mathbf{y}$, where $\mathbf{Z}$ is a matrix of instruments conformable to the matrix of regressors, $\mathbf{W}$, with rows including $X_1$ and $S$, and $\mathbf{y}$ is a vector of observations on the dependent variable. Because $X_1$ is assumed to be uncorrelated with $\varepsilon$, $\mathbf{Z}$ typically includes $X_1$ and a single variable not in $X_1$, perhaps taken from $X_2$ in (2). For example, $X_2$ might include quarter of birth. For IV to be consistent, $\mathbf{Z}$ must be asymptotically uncorrelated with the regression error and the probability limit of $(\mathbf{Z}'\mathbf{W}/n)$ must be nonsingular ($n$ is the sample size).

An alternative estimation strategy based on the same idea is TSLS. There are potentially three different IV or Wald estimates that could be computed using quarter-of-birth dummies. A TSLS estimator using all the information available on quarter of birth is calculated by first regressing $S$ (the endogenous regressor) on all covariates included in the equation, $X_1$, and on all the potential instruments excluded from the equation, $X_2$—in this case three quarter-of-birth dummies. The second stage in the TSLS procedure is to estimate

$$Y = \gamma_0 + X_1\gamma_1 + \rho\hat{s} + \nu,$$

where $\hat{s}$ is the fitted value from the first-stage regression and $\nu \equiv \{\varepsilon + \rho[S - \hat{s}]\}$. In this case TSLS can be interpreted as an IV estimator where the instruments are $X_1$ and $\hat{s}$ (Theil 1971), or as an efficient linear combination of alternative IV estimates using single quarter-of-birth dummies.

Columns (1)–(2) and (5)–(6) of Table 2 report OLS and TSLS estimates of Equation (1) without covariates using both Census data sets. The excluded instruments used to construct the TSLS estimates in these columns are three quarter-of-birth dummies. That is, the estimates are regressions of the log weekly wage on fitted values from a "first-stage" regression of schooling on a constant and three quarter-of-birth dummies.

Columns (3)–(4) and (7)–(8) report estimates based on equations where the set of covariates, $X_1$ includes nine year-of-birth dummies. The excluded instruments used to construct the TSLS estimates in columns (4) and (8) include three quarter-of-birth dummies interacted with 10 year-of-birth dummies. This specification allows for year-of-birth effects in the outcome equation and allows the relationship between schooling and quarter of birth to differ for each year of birth. The OLS and TSLS estimates are similar and differ little across model specifications.

## 3. THE AVERAGE CAUSAL EFFECT OF A VARIABLE TREATMENT

Equation (1) is a structural relationship derived from assumptions about human behavior, but it is not necessarily a causal relationship in the Rubin (1974) sense. In this section we present results that give IV and TSLS estimates of Equation (1) a causal interpretation. Suppose that each individual would earn $Y_j$ if he or she had $j$ years of schooling for $j = 0, 1, 2, \ldots, J$. It is useful to imagine that a full set of $Y_j$ exists for each person, even though only one is observed. The set of $Y_j$ for one person is assumed independent of the outcomes and treatment status of other people. Rubin has called independence of these potential or counterfactual outcomes across individuals the stable unit treatment values (SUTVA) assumption. A full set of counterfactual outcomes and SUTVA are commonly assumed in the statistics literature on causality (e.g., Holland 1986), although these are not trivial assumptions. Elsewhere (Angrist et al. 1995), we have discussed conceptual issues associated with these assumptions in an IV context.

Provided that we are willing to entertain the notion of counterfactual outcomes, the objective of causal inference is to uncover information about the distribution of $Y_j - Y_{j-1}$, which is the causal effect of the $j$th year of schooling (Holland 1986; Rubin 1974, 1977). We view estimates of $\rho$ in Equation (1) as having a causal interpretation when they have probability limit equal to a weighted average of $E[Y_j - Y_{j-1}]$ for all $j$ in some subpopulation or subpopulations of interest. In general, this will not be the case. (If the treatment level, $S$, is randomly assigned, then $E[Y_j - Y_{j-1}]$ can be consistently

*Table 2. TSLS Estimates of the Return to Schooling*

| | Born 1920–1929, 1970 census | | | | Born 1930–1939, 1980 census | | | |
|---|---|---|---|---|---|---|---|---|
| | OLS | TSLS | OLS | TSLS | OLS | TSLS | OLS | TSLS |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Education | .080<br>(.0004) | .063<br>(.016) | .080<br>(.0004) | .077<br>(.015) | .071<br>(.0003) | .103<br>(.020) | .071<br>(.0003) | .089<br>(.016) |
| YOB dummies | no | no | yes | yes | no | no | yes | yes |
| $\chi^2$(dof) | | 2.4 (2) | | 36.0 (29) | | 2.9 (2) | | 25.4 (29) |
| Sample size | | 247,199 | | | | 329,509 | | |

NOTE: Each sample consists of men born in any quarter in the United States who had positive earnings in the year preceding the survey. The table reports coefficients from OLS and TSLS regressions of the log weekly wage on years of completed schooling. The excluded instruments in columns 2 and 6 are three quarter-of-birth dummies. The excluded instruments in columns 4 and 8 are three quarter-of-birth dummies times 10 year-of-birth dummies. The $\chi^2$ statistic is an instrument-error orthogonality test statistic.

estimated by subtracting the average response for individuals with treatment level $j-1$ from the average response for individuals with treatment level $j$.)

We define $S_Z \in \{0, 1, 2, \ldots, J\}$ to be the number of years of schooling completed by a student conditional on the student's quarter of birth, $Z$. As with $Y_j$, $S_Z$ is assumed to exist for each value of $Z$ for each person, even though only one $S_Z$ is observed. This setup incorporates two innovations to the original Rubin (1974) framework by allowing for both multivalued treatments and counterfactual treatment status. Rubin (1978) and Robins (1989a, b) also discussed causal inference with multivalued treatments, and Robins (1989a), Robins and Greenland (1992) and Holland (1988) used the notion of counterfactual treatment status. As far as we know, however, these ideas have not been used previously in an IV or TSLS framework.

Initially we assume that $Z$ is coded to take on only two values, 0 and 1, indicating first or later quarters of birth. $S_0$ is the years of schooling that would be attained by an individual born in the first quarter, and $S_1$ is the years of schooling that would be attained by the same individual if he or she were to be born in later quarters. For each person in the Census sample, we observe the triple $(Z, S, Y)$, where $Z$ is the quarter of birth, $S = S_Z = Z \cdot S_1 + (1 - Z) \cdot S_0$ is years of completed schooling, and $Y = Y_S$ is earnings. Our principal identifying assumption (apart from assuming the existence of $Y_j$) is that $Z$ is independent of all potential outcomes and potential treatment intensities. As a practical matter, this assumption may be true only after conditioning on covariates. (In fact, the need to control for covariates sometimes motivates the use of TSLS instead of IV.) Formally, we have the following assumption.

*Assumption 1 (Independence).* The random variables $S_0$, $S_1$, $Y_0$, $Y_1, \ldots, Y_J$ are jointly independent of $Z$.

In the compulsory schooling example, Assumption 1 requires that quarter of birth has no effect on earnings other than through its effect on schooling. This is the meaning of the notion that quarter of birth provides a natural experiment that can be used to estimate the effect of schooling on earnings. Assumption 1 can also be viewed as a nonparametric version of the assumptions that the instruments, $X_2$, are uncorrelated with $\varepsilon$ and $\eta$, the error terms in Equations (1) and (2). Angrist et al. (1995) discussed the impact of deviations from Assumption 1 on IV estimates in the binary treatment case.

It is important to note that outside a linear regression framework, the independence assumption alone is not usually sufficient to identify a meaningful average treatment effect. This point is easiest to see in a simple example where $S$ is binary (perhaps indicating high school graduates versus nongraduates). A comparison of outcomes by different values of the instrument gives

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] \\ = E[Y_0 + (Y_1 - Y_0)S_1 | Z = 1] \\ - E[Y_0 + (Y_1 - Y_0)S_0 | Z = 0]. \end{aligned} \quad (3)$$

By Assumption 1, this simplifies to

$$\begin{aligned} E[Y_0 + (Y_1 - Y_0)S_1] - E[Y_0 + (Y_1 - Y_0)S_0] \\ = E[(Y_1 - Y_0)\cdot(S_1 - S_0)]. \end{aligned} \quad (4)$$

Without imposing additional restrictions, $S_1 - S_0$ can be equal to 1, 0, or $-1$. A value of 1 indicates individuals who are induced to graduate high school by the instrument (i.e., being born in a late quarter), a value of 0 indicates those whose schooling status is unchanged, and a value of $-1$ indicates those who are induced to drop out of high school before graduating. Therefore,

$$\begin{aligned} E[(Y_1 - Y_0)\cdot(S_1 - S_0)] \\ = E[Y_1 - Y_0 | S_1 - S_0 = 1]\cdot \Pr[S_1 - S_0 = 1] \\ - E[Y_1 - Y_0 | S_1 - S_0 = -1]\cdot \Pr[S_1 - S_0 = -1]. \end{aligned} \quad (5)$$

Individuals whose schooling status is unaffected by the instrument clearly do not contribute anything to comparisons of average outcomes by instrument status. But the group that does contribute includes both "switchers-in" and "switchers-out." It is clear that it is theoretically possible to have a situation where the treatment effect, $Y_1 - Y_0$, is positive for everyone but the sizes of the group of switchers-in and switchers-out is such that the average difference in outcomes is zero or even negative. Suppose, for example, that the treatment effect (effect of graduating high school) equals $\alpha$ for those induced to graduate and $2\alpha$ for those induced to drop out. If $\Pr[S_1 - S_0 = 1] = \frac{2}{3}$ and $\Pr[S_1 - S_0 = -1] = \frac{1}{3}$, then the average difference in $Y$ conditional on $Z$ is zero, even though $Y_1 - Y_0 > 0$ for everyone.

The most common way to get around this problem is simply to assume a constant unit treatment effect, $Y_j - Y_{j-1} = \alpha$, for all $j$ and all individuals. This is the assumption underlying econometric applications using linear regression models, as well as the application of instrumental variables techniques by Permutt and Hebel (1989). In his comment on Holland's (1988) discussion of causality, Leamer (1988) pointed out that in linear models with constant treatment effects, the problem of using instrumental variables for causal inference is straightforward.

Instead of restricting treatment effect heterogeneity, in this article we impose a nonparametric restriction on the process determining $S$ as a function of $Z$. This restriction is that either $S_1 - S_0 \geq 0$ or $S_1 - S_0 \leq 0$ for everyone. If in the binary treatment example, $S_1 - S_0 \geq 0$, then (5) becomes

$$\begin{aligned} E[(Y_1 - Y_0)\cdot(S_1 - S_0)] \\ = E[Y_1 - Y_0 | S_1 - S_0 = 1]\cdot \Pr[S_1 - S_0 = 1]. \end{aligned} \quad (6)$$

The conditional expectation, $E[Y_1 - Y_0 | S_1 - S_0 = 1]$, is what we have called a local average treatment effect (LATE; Imbens and Angrist 1994). LATE is the average causal effect of treatment for those whose treatment status is affected by the instrument (i.e., for those for whom $S_1 = 1$ and $S_0 = 0$). Formally, the following monotonicity condition is sufficient (given independence) for LATE to be identified.

*Assumption 2 (Monotonicity).* With probability 1, either $S_1 - S_0 \geq 0$ or $S_1 - S_0 \leq 0$ for each person.

This assumption has been discussed previously by Robins (1989a), who showed that it does not sharpen bounds for population-average treatment effects, and by Permutt and Hebel (1989) in the context of a regression model for the effect of maternal smoking on birth weight. In the smoking example, monotonicity means that a randomized antismoking intervention never increases smoking. In the schooling example, monotonicity means that, because of compulsory attendance laws, people born in quarters 2–4 complete at least as much schooling as they would have completed had they been born in the first quarter.

Assumption 2 is not verifiable, because it involves unobserved variables (i.e., only one of $S_1$ or $S_0$ is observed). Nevertheless, for multivalued treatments ($J > 1$), Assumption 2 has the testable implication that the cumulative distribution function (CDF) of $S$ given $Z = 1$ and the CDF of $S$ given $Z = 0$ should not cross, because if $S_1 \geq S_0$ with probability 1, then $\Pr(S_1 \geq j) \geq \Pr(S_0 \geq j)$ for all $j$. This implies $\Pr(S \geq j | Z = 1) \geq \Pr(S \geq j | Z = 0)$ or $F_S(j | Z = 0) \geq F_S(j | Z = 1)$, where $F_S$ is the CDF of $S$. We investigate this implication in the schooling example that follows. If $J = 1$ and the treatment is binary, then the CDF's cannot cross.

The main theoretical result of the article is now given for the case where $S_1 - S_0 \geq 0$.

*Theorem 1.* Suppose that Assumptions 1 and 2 hold and that $\Pr(S_1 \geq j > S_0) > 0$ for at least one $j$. Then

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[S|Z=1] - E[S|Z=0]} = \sum_{j=1}^{J} \omega_j \cdot E[Y_j - Y_{j-1} | S_1 \geq j > S_0] \equiv \beta, \quad (7)$$

where

$$\omega_j = \frac{\Pr(S_1 \geq j > S_0)}{\sum_{i=1}^{J} \Pr(S_1 \geq i > S_0)}.$$

This implies that $0 \leq \omega_j \leq 1$ and $\sum_{j=1}^{J} \omega_j = 1$, so that $\beta$ is a weighted average per-unit treatment effect.

Proof of the theorem is given in the Appendix. The proof follows the same lines as the development from Equations (3)–(6) and generalizes our earlier result (Imbens and Angrist 1994) to models with variable treatment intensity. A simplified example with three treatment intensities helps to understand the more general result. Suppose that $S$ can be equal to 0, 1, or 2. We can write

$$Y = Y_0 + (Y_1 - Y_0) I[S \geq 1] + (Y_2 - Y_1) I[S \geq 2],$$

where $I[A]$ is the indicator function for event $A$. A version of Equation (4) for this case is

$$\begin{aligned} E[Y|Z=1] - E[Y|Z=0] &= E[(Y_1 - Y_0) \cdot (I[S_1 \geq 1] - I[S_0 \geq 1])] \\ &\quad + E[(Y_2 - Y_1) \cdot (I[S_1 \geq 2] - I[S_0 \geq 2])]. \end{aligned}$$

By virtue of Assumption 2, $I[S_1 \geq 1] - I[S_0 \geq 1]$ and $I[S_1 \geq 2] - I[S_0 \geq 2]$ must both be either 1 or 0. Therefore, $\Pr\{I[S_1 \geq 1] - I[S_0 \geq 1] = 1\} = \Pr(S_1 \geq 1 > S_0)$ and $\Pr\{I[S_1 \geq 2] - I[S_0 \geq 2] = 1\} = \Pr(S_1 \geq 2 > S_0)$.

The requirement that $\Pr(S_1 \geq j > S_0) > 0$ for some $j$ means that the instrument must affect the level of treatment, $S$. Also, note that in the proof of Theorem 1, $S$ is assumed to take on only integer values between 0 and $J$. It is enough, however, that $S$ be bounded and take on a finite number of rational values. Then one can always use a linear transformation to ensure that $S$ takes on integer values only between 0 and $J$. A linear transformation of $S$ does not have any effect on the numerator of (7) and multiplies the denominator by a constant. Thus the linear transformation amounts to changing the units in which treatment intensity is measured.

Theorem 1 is important because it shows that in a wide variety of models and circumstances, it is possible to identify features of the distribution of $Y_j - Y_{j-1}$. For example, the monotonicity assumption appears plausible in research designs based on the draft lottery (e.g., Angrist 1990) and in designs based on randomly assigned encouragement or intention-to-treat such as discussed by Powers and Swinton (1984) and Holland (1988). This assumption is also mechanically satisfied in the latent index models commonly used in econometrics (Imbens and Angrist 1994).

We refer to the parameter $\beta$ as the average causal response (ACR). This parameter captures a weighted average of causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. The weight attached to the average of $Y_j - Y_{j-1}$ is proportional to the number of people who, because of the instrument, change their treatment from less than $j$ units to $j$ or more units. This proportion is $\Pr(S_1 \geq j > S_0)$. In the schooling example, this is the proportion of people who, by accident of birth, are induced to complete additional years or fractional years of schooling. Note that this group need not be representative of the population, and that the members of this group cannot be identified from the data because membership involves unobserved counterfactual treatment status.

A referee made the point that although the ACR is a weighted average, it averages together components that are potentially overlapping. For example, someone who is induced to graduate high school by having been born in a late quarter, but would have completed only 11th grade had he or she been born in the first quarter, contributes to the population of individuals for whom $\Pr(S_1 \geq 12 > S_0)$. But anyone who is induced to graduate high school, but would have otherwise completed only 10th grade, will be in the population of individuals for whom $\Pr(S_1 \geq 12 > S_0)$ *and* for whom $\Pr(S_1 \geq 11 > S_0)$.

Similarly, suppose that the instrument induces some fraction of the sample to go from 10 to 12 units of treatment but has no effect otherwise. Then the ACR can be written as the sum of two single-unit average effects,

$$\frac{1}{2} E[Y_{12} - Y_{11} | S_1 \geq 12 > S_0] + \frac{1}{2} E[Y_{11} - Y_{10} | S_1 \geq 11 > S_0],$$

although what really is identified is $E[Y_{12} - Y_{10} | S_1 \geq 12, 11 > S_0]$. In the schooling and other examples, however, most individuals would probably not be involved in an overlap of this sort, because the instrument would typically be expected to cause no more than a one-unit increment in treatment intensity for any particular individual.

### 3.1 Incorrectly Coded Binary Treatments

Theorem 1 has a simple corollary that can be used to interpret parameter estimates in models where a variable treatment is incorrectly parameterized as a binary treatment. For example, Permutt and Hebel (1989) discussed conditions sufficient to identify the effect of smoking when it is assumed that all that matters for health is whether any cigarettes are smoked. Similarly, econometricians sometimes estimate the effect of college and/or high school graduation on earnings, ignoring the fact that dummy variables indicating graduation are nonlinear functions of an underlying years-of-schooling variable (e.g., Rosen and Willis 1979).

*Corollary (Misspecified binary treatment).* Suppose that the treatment of interest is assumed to be an indicator function of $S$, say $b \equiv I(S \geq l)$, for some $1 \leq l \leq J$. Then, given Assumptions 1 and 2,

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[b|Z=1] - E[b|Z=0]} = \phi \cdot \beta,$$

where

$$\phi \equiv \frac{E[S|Z=1] - E[S|Z=0]}{E[b|Z=1] - E[b|Z=0]} = \frac{\sum_{j=1}^{J} \Pr(S_1 \geq j > S_0)}{\Pr(S_1 \geq l > S_0)} \geq 1.$$

Note that the only situation where $\phi = 1$ is when the instrument has no effect other than to cause people to switch from $S = l - 1$ to $S = l$. Thus when a variable treatment is incorrectly parameterized as binary, the resulting estimate tends to be too large relative to the average per-unit effect along the length of the response function. On the other hand, by virtue of monotonicity, the sign of the ACR is still consistently estimated.

### 3.2 Estimation of the ACR and the Weighting Function

A natural estimator of $\beta$ is its sample analog. This estimator is an application of Wald's (1940) grouping method of fitting straight lines, where the data have been grouped by the instrument. Durbin (1954) appears to have been the first to point out that the Wald estimator is also an instrumental variables estimator.

The ACR weights in Theorem 1 can be estimated using a random sample of $(Y, S, Z)$ because

$$\begin{aligned}\Pr(S_1 \geq j > S_0) &= \Pr(S_1 \geq j) - \Pr(S_0 \geq j) \\ &= \Pr(S_0 < j) - \Pr(S_1 < j) \\ &= \Pr(S < j | Z = 0) - \Pr(S < j | Z = 1).\end{aligned}$$

Thus the weighting function can be consistently estimated from the difference between the empirical CDF's of $S$ given $Z$.

## 4. MULTIPLE INSTRUMENTS AND MODELS WITH COVARIATES

Because different instruments are associated with different weighting schemes in the definition of the ACR, the discussion in the previous section provides an explanation of why estimates of $\beta$ constructed using different instruments might differ. The typical econometric application of TSLS, however, imposes a constant-treatment-effect model in which $Y_j - Y_{j-1} = \alpha$ for all $j$ and all individuals. In this case, alternative instrumental variables estimates of the same $\alpha$ can be combined into a single, more efficient estimate using TSLS. What does the TSLS estimator—which combines alternative instrumental variables estimates—produce when it is applied to the heterogeneous-treatment-effects model outlined in Section 3?

We explore this question for the case where $K$ mutually orthogonal binary instruments are combined to form a single TSLS estimate. This is a fairly general example, because any set of discrete instruments can be recoded into a set of mutually orthogonal indicator variables. TSLS using $K$ orthogonal indicators can be thought of as a means of exploiting a single $K + 1$-valued instrument, $Z$ (e.g., quarter of birth takes on four possible values). Moreover, a saturated model for the first stage consistently estimates the conditional expectation of the endogenous regressors given the instrument. This leads to the most efficient TSLS estimator in homoscedastic regression models with constant treatment effects (Newey 1990).

Theorem 2 shows that the TSLS estimator constructed by using a constant plus $K$ linearly independent dummy variables, $d_k = I(Z = k)$, as instruments has probability limit equal to a weighted average of $K$ linearly independent ACR's, $\beta_{k,k-1}$, where

$$\beta_{k,k-1} \equiv \frac{E[Y|Z=k] - E[Y|Z=k-1]}{E[S|Z=k] - E[S|Z=k-1]}.$$

Because each $\beta_{k,k-1}$ is a weighted average of points on the causal response function, the TSLS estimate also converges to a weighted average of points on the causal response function.

Let the points of support of $Z$ be ordered such that $l < m$ implies $E[S|Z = l] < E[S|Z = m]$. Note that using $K$ dummies, $d_k = I(Z = k)$, plus a constant in TSLS estimation is the same as instrumental variables estimation using $E[S|Z]$ plus a constant as instruments. We then have the following theorem.

*Theorem 2.* Suppose that $E[S|Z]$ and a constant are used to construct instrumental variables estimates of $\beta_z$ in the equation

$$Y = \gamma + \beta_z S + \varepsilon.$$

The resulting estimate has probability limit

$$\beta_z \equiv \frac{E\{Y \cdot (E[S|Z] - E[S])\}}{E\{E[S|Z] \cdot (E[S|Z] - E[S])\}} = \sum_{k=1}^{K} \mu_k \beta_{k,k-1},$$

where

$$\mu_k = (E[S|Z=k] - E[S|Z=k-1]) \cdot \frac{\sum_{l=k}^{K} \pi_l(E[S|Z=l] - E[S])}{\sum_{l=0}^{K} \pi_l E[S|Z=l](E[S|Z=l] - E[S])}$$

and $\pi_l = \Pr[Z = l]$. Moreover, $0 \le \mu_k \le 1$ and $\sum_{k=1}^{K} \mu_k = 1$.

This result modifies and generalizes a previous result of ours (Imbens and Angrist 1994) for binary treatments. Again, details of the proof are in the Appendix. The theorem follows partly from standard formulas interpreting TSLS using mutually orthogonal instruments as a weighted average of each of the instrumental variables estimates obtained taking the instruments one by one. Moreover, when the instruments are mutually exclusive dummy variables, TSLS can be written as a linear combination of linearly independent Wald estimates (Angrist 1991). The proof essentially combines these two results.

Theorem 2 provides a useful interpretation for conventional TSLS estimates. Just as the simple Wald estimator converges to a weighted average effect along the length of the causal response function, TSLS estimates provide one way of combining a set of different weighted average effects into a new weighted average. The weights used to construct TSLS estimates from Wald estimates are proportional to $(E[S|Z=k] - E[S|Z=k-1])$. Thus the better the Wald estimate, in the sense of being based on an instrument with a bigger impact on the regressor, the more weight it receives in the TSLS linear combination. The second component of the weighting function, $\sum_{l=k}^{K} \pi_l(E[S|Z=l] - E[S])$, simplifies to $[E(S|Z \ge k) - E(S|Z < k)]P(Z \ge k)[1 - P(Z \ge k)]$. Thus TSLS gives more weight to Wald estimates that are closer to the center of the distribution of $Z$.

### 4.1 TSLS Estimates of Models with Covariates

Conditional on discrete covariates such as year of birth, the problem of identifying and estimating the ACR is exactly the same as outlined previously. Therefore, analysis of models with discrete covariates can proceed in subsamples where the covariates are fixed. A more parsimonious approach exploits the fact that instrumental variables estimates of average treatment effects have a useful averaging property in pooled subsamples. In particular, ignoring the fact that the ACR may vary with the covariates leads to a variance-weighted average treatment effect.

The following result formally describes the probability limit of the instrumental variables estimator when we allow for a changing intercept but fix the treatment effect across covariates.

*Theorem 3.* Let $\mathbf{g}[X]$ be a design matrix constructed from indicator variables for each value of $X$. Consider the TSLS estimate computed using $\mathbf{g}(X)$ and a full set of interactions between $\mathbf{g}(X)$ and $Z$ as instruments for a regression of $Y$ on rows of $\mathbf{g}[X]$ and $S$. The resulting estimate is

$$\beta_x = \frac{E\{Y \cdot (E[S|X,Z] - E[S|X])\}}{E\{S \cdot (E[S|X,Z] - E[S|X])\}} \tag{8}$$

$$= \frac{E\{\beta(X)\Theta(X)\}}{E[\Theta(X)]}, \tag{9}$$

where $\Theta(X) = E\{E[S|X,Z] \cdot (E[S|X,Z] - E[S|X])|X\}$ and

$$\beta(X) = \frac{E\{Y \cdot (E[S|X,Z] - E[S|X])|X\}}{E\{S \cdot (E[S|X,Z] - E[S|X])|X\}}. \tag{10}$$

*Proof.* Equation (8) is immediate from the definition of TSLS using dummy variable instruments. The weighting formulas (9) and (10) can be established by iterating expectations and using the definition of $\beta_z$ from Theorem 2.

Note that $\beta(X)$ is the TSLS estimate, $\beta_z$, constructed using $Z$ as an instrument in a population where $X$ is fixed. Thus Theorem 3 says that the TSLS estimates of a single treatment effect in a model with dummy variable covariates is a weighted average of the TSLS estimates conditional on the covariates. The weights consist of the variance of $E[S|X,Z]$ conditional on the covariates.

### 4.2 Inference

In another paper (Imbens and Angrist 1994), we discussed results on the asymptotic variance of IV and TSLS estimates in models with binary treatments. These results apply to the case discussed here, and they imply that standard errors of the ACR can be calculated using formulas of Huber (1967) and White (1982). One reason for reporting TSLS estimates as well as Wald estimates is that in models with constant treatment effects, the TSLS estimates have asymptotically lower sampling variance than any single Wald estimate. In general, however, this need not be true if there is variation in the average causal response across instruments. Nevertheless, TSLS provides a convenient way to combine alternative IV estimates in a single statistic.

TSLS estimators are also associated with an overidentification test statistic that equals the objective function implicitly minimized by the estimates (Newey 1985). In a constant-treatment–effect model estimated by TSLS, the statistic provides an overidentification test for the null hypothesis that all the instruments are orthogonal to the regression error term. The constant treatment effect is overidentified because any single instrument would be sufficient for identification. But in the model outlined here, each instrument can lead to a different estimate even though all the instruments satisfy the independence assumption. In fact, Theorems 1, 2, and 3 provide possible explanations for why estimates of causal effects such as the economic returns to schooling may differ in studies using different samples or in a single sample with different instruments and covariates. For example, a recent study using instrumental variables to estimate the returns to schooling in a sample of twins (Ashenfelter and Krueger 1994) leads to estimated coefficients roughly double those reported here and by Angrist and Krueger (1991).

## 5. IV ESTIMATES OF THE RETURNS TO SCHOOLING: FOR WHOM?

Angrist and Krueger (1991) used linear regression models with constant coefficients to interpret estimates of the return to schooling based on quarter of birth. In the context of the causal model outlined here, however, the Wald estimates in

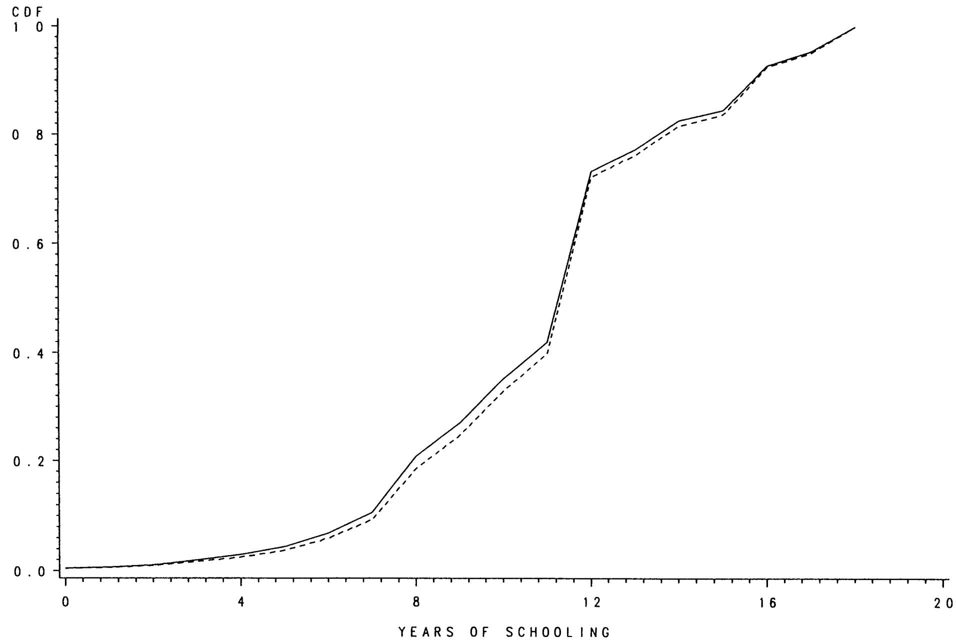


*Figure 1. Schooling CDF by Quarter of Birth (Men Born 1920–1929; Data From the 1970 Census). Quarter of birth: ——, first; - - -, fourth.*

Table 1 should be interpreted as the average effect of a 1-year increase in schooling, for people whose schooling is influenced by quarter of birth. This is a small group, not necessarily representative of the entire population. To identify the ACR for this group, the monotonicity condition requires that men born in the fourth quarter get at least as much schooling as they would have if they had been born in the first quarter. If this condition is satisfied, then we can get some idea of the size and characteristics of the group contributing to the ACR through the ACR weighting function.

The CDF's of schooling by quarter of birth for men in the 1970 and 1980 Censuses are graphed in Figures 1 and 2. Both figures show that the CDF for men born in the fourth quarter lies below the CDF for men born in the first quarter. This is important evidence in favor of the monotonicity assumption in this example. The weighting function underlying estimates of the ACR in Table 2 is proportional to the difference between the CDF of schooling for men born in the first quarter and the CDF of schooling for men born in the fourth quarter. For each level of schooling, $j$, this difference is the fraction of the population whose schooling is switched by quarter of birth from less than $j$ years to at least $j$ years.

Figures 3 and 4 show differences in the CDF of schooling by quarter of birth. In each figure, differences between the

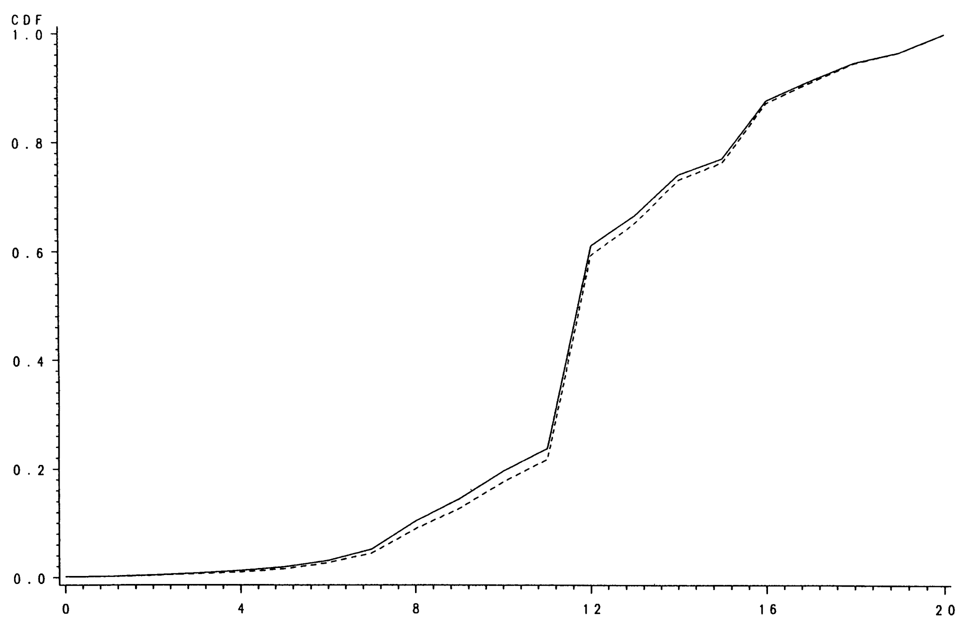


*Figure 2. Schooling CDF by Quarter of Birth (Men Born 1930–1939; Data From the 1980 Census). Quarter of birth: ——, first; - - -, fourth.*

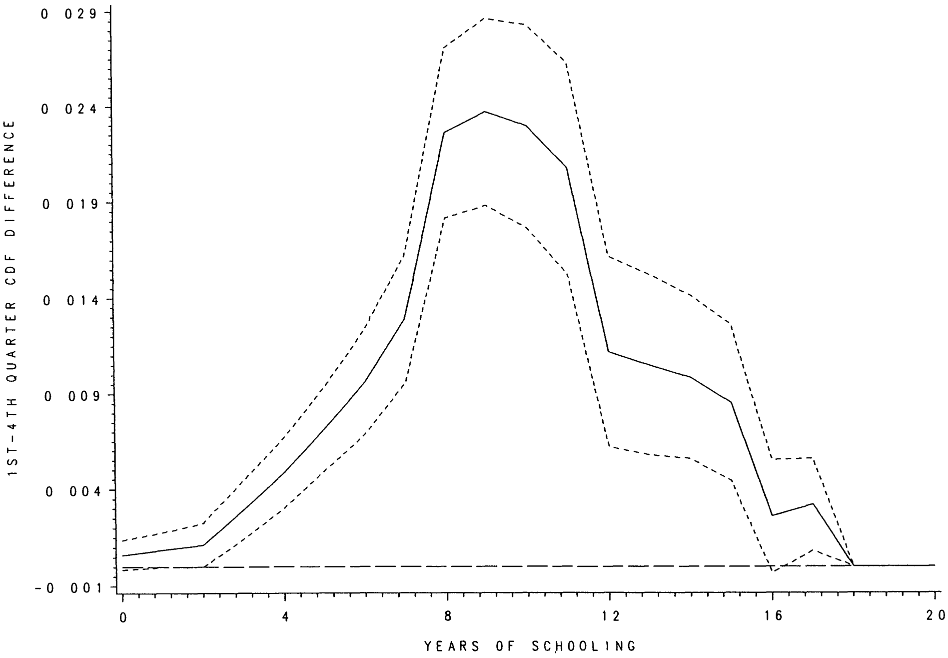


*Figure 3. First-Fourth Quarter Difference in Schooling CDF (Men Born 1920-1929, Data From the 1970 Census). Dotted lines are 95% confidence intervals.*

CDF of schooling for men born in the first and fourth quarters are plotted, along with 95% pointwise confidence bands (calculated using the conventional formula for a difference in proportions). ACR weighting functions for estimates based on comparisons between first- and fourth-quarter births are the CDF differences plotted in the figures, normalized to sum to 1.

The figures show that the groups contributing most to estimates of the ACR based on quarter of birth are those with 8–12 years of schooling. Both figures show declines in the weighting function at around 12 years of schooling. A maximum of around 2% of the sample was induced by being born in the fourth quarter to complete 11th grade, but much smaller fractions were induced to complete higher grades. This is not surprising, because compulsory attendance laws affect mainly high school students and cannot compel students to go to college. Note that some weight is contributed by college attenders, perhaps because some students forced

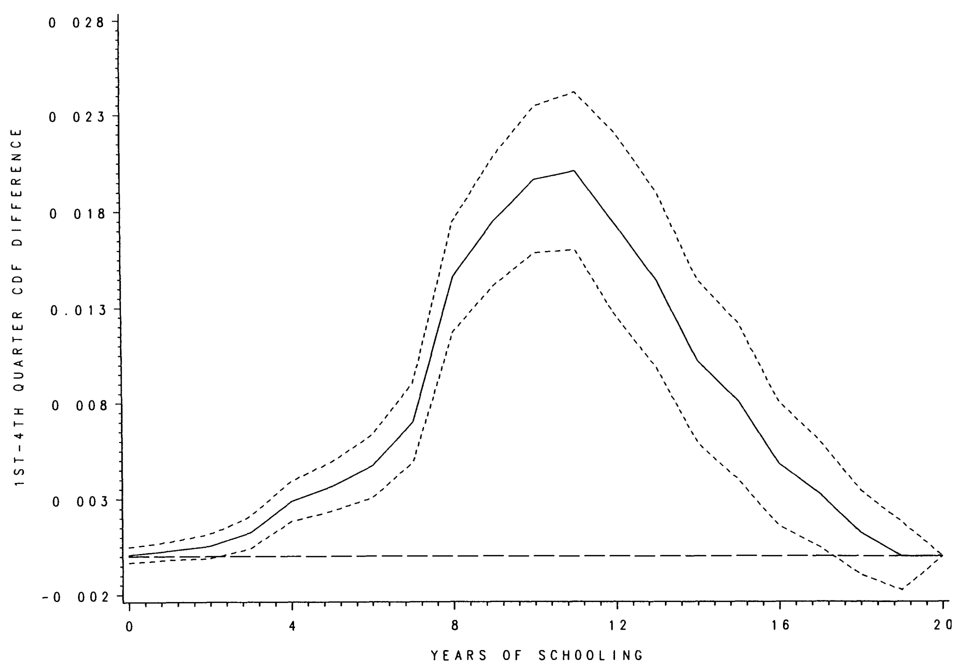


*Figure 4. First-Fourth Quarter Difference in Schooling CDF (Men Born 1930-1939; Data From the 1980 Census). Dotted lines are 95% confidence intervals.*

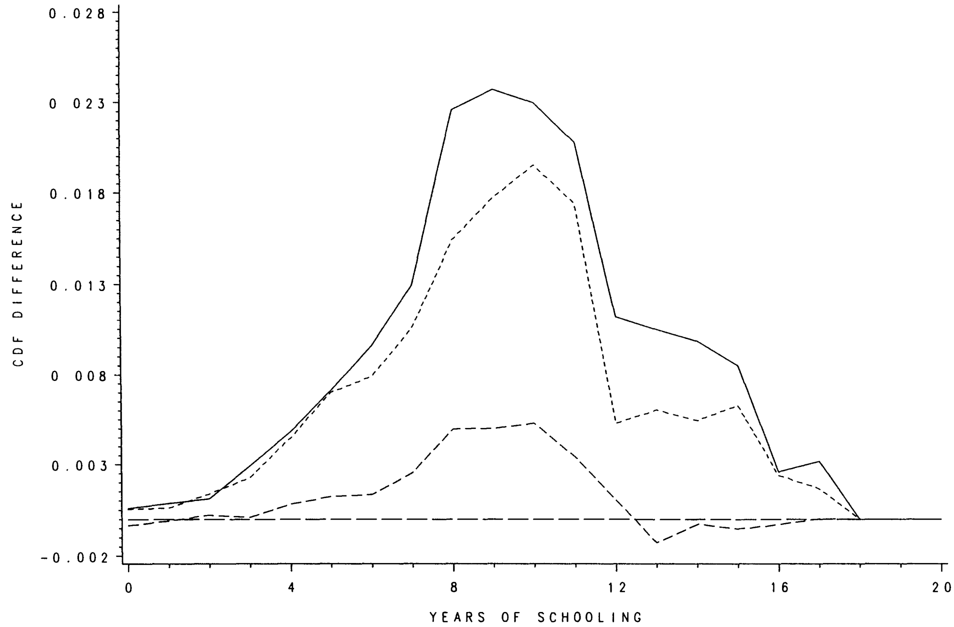


*Figure 5. Differences in Schooling CDF by Quarter of Birth (Men Born 1920–1929; Data From the 1970 Census). Quarter of birth: ——, 1st–4th; - - -; 2nd–4th; — —, 3rd–4th.*

by accident of birth to graduate high school decided later to go on to college after all.

One interesting feature of Figures 3 and 4 is that Figure 3, for men born in 1920–1929, shows a much sharper drop at 12 years of schooling than does Figure 4, for men born in 1930–1939. Therefore, men who ended up completing some college because they were forced to graduate high school contribute more to the estimates for men born in 1930–1939 than to the estimates for men born in 1920–1929. This difference may explain the higher Wald and TSLS estimates for men born in 1930–1939 (despite the fact that OLS estimates for the more recent cohort are lower), because the returns to the last year of college tend to be substantially higher than those for any single year of high school (Card and Krueger 1992).

Figures 5 and 6 plot the contrast between schooling CDF's for birth quarters 1–3 relative to fourth-quarter births. The figures show that schooling CDF's are essentially ordered by quarter of birth. This is evidence that any adjacent pair of quarters can be used to define a binary instrumental variable that satisfies the monotonicity assumption. TSLS using three quarter-of-birth dummies is a weighted average of the three

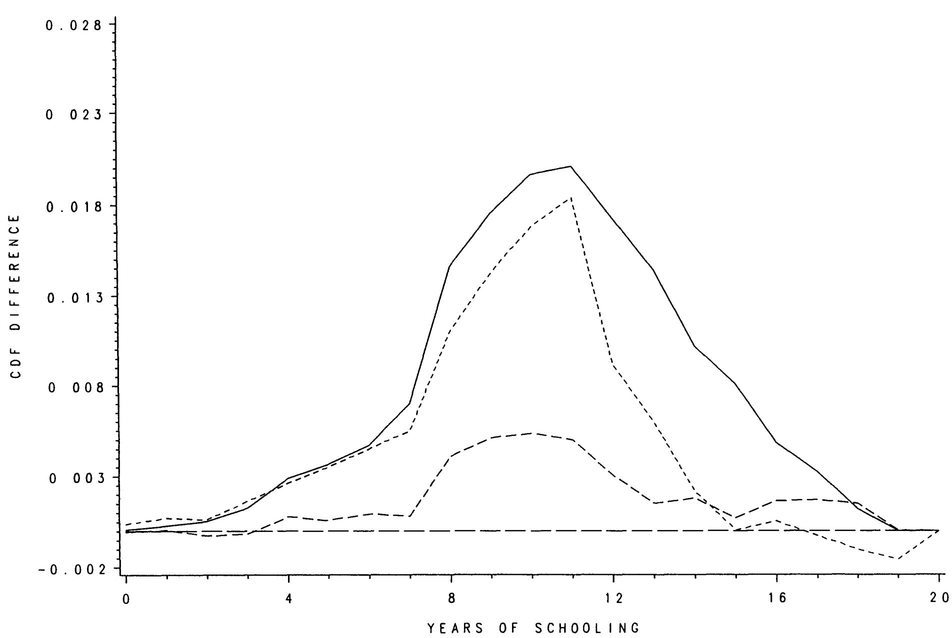


*Figure 6. Differences in Schooling CDF by Quarter of Birth (Men Born 1930–1939; Data From the 1980 Census). Quarter of birth: ——, 1st–4th; - - -, 2nd–4th; — —, 3rd–4th.*

possible Wald estimates based on adjacent quarters of birth. The TSLS estimates, reported in Table 2, are .063 in the 1970 Census and .103 in the 1980 Census. These are estimated with slightly greater precision than the Wald estimates reported in Table 1.

Estimates of models including year of birth dummies are reported in columns (3)–(4) and (7)–(8) of Table 2. The instrument list for these models includes a set of three quarter-of-birth dummies for each year of birth. In the context of Theorem 3, the TSLS estimates in columns (4) and (8) can be interpreted as a weighted average of separate TSLS estimates of the ACR for each year of birth.

The TSLS overidentification test statistics for each of the models reported in Table 2 are far from critical values at conventional significance levels under the null hypothesis of constant treatment effects and instrument error orthogonality. Thus the test statistics cast little doubt on the constant-treatment–effect and independence assumptions.

## 6. SUMMARY AND CONCLUSIONS

This article defines the average causal response to variable treatments such as drug dosage, cigarettes smoked, hours of study, and years of schooling. We have shown here that a weighted average of per-unit causal responses to a change in treatment intensity is identified in a wide variety of models and circumstances. The average response that we can identify is for individuals whose treatment status is affected by an instrumental variable that is independent of potential outcomes and potential treatment intensities. The monotonicity condition imposed when deriving this result requires only that the instrumental variable affect treatment intensity in the same direction for each unit of observation. This condition has testable implications in models with variable treatment intensities.

We have presented formulas for the weighting functions that underlie IV and TSLS estimates of average causal effects. These formulas can help empirical researchers understand which observations are contributing to a particular estimate and provide a causal interpretation for some of the simple estimators commonly used in applied research. The interpretation of TSLS and the example presented here serve to emphasize a point made earlier by Rubin (1986), that observational data can only be informative about the causal effect of treatment for those whose treatment status can be thought of as having been manipulated in some way. This paper shows that the estimated treatment effect may change when the nature of this manipulation changes.

## APPENDIX: PROOFS

### Proof of Theorem 1

Let $I(A)$ be the indicator function for the event $A$. Define the following indicators: $\lambda_{Zj} = I(S_Z \geq j)$ for $Z = 0$, 1 and $j = 0, 1, 2, \ldots, J+1$. Note that $\lambda_{Z0} = 1$ and $\lambda_{ZJ+1} = 0$ for all $Z$. In terms of the $\lambda_{Zj}$, $Y$ can be written as

$$Y = Z \cdot Y_{S_1} + (1-Z) \cdot Y_{S_0}$$

$$= \left\{ Z \cdot \sum_{j=0}^{J} Y_j(\lambda_{1j} - \lambda_{1j+1}) \right\} + \left\{ (1-Z) \cdot \sum_{j=0}^{J} Y_j(\lambda_{0j} - \lambda_{0j+1}) \right\}.$$

Using the independence assumption, $E[Y|Z=1] - E[Y|Z=0]$ is, therefore,

$$E\left\{ \sum_{j=0}^{J} Y_j \cdot [\lambda_{1j} - \lambda_{1j+1} - \lambda_{0j} + \lambda_{0j+1}] \right\}$$

$$= E\left\{ \sum_{j=1}^{J} [(Y_j - Y_{j-1}) \cdot (\lambda_{1j} - \lambda_{0j})] + Y_0 \cdot (\lambda_{10} - \lambda_{00}) \right\}$$

$$= E\left\{ \sum_{j=1}^{J} (Y_j - Y_{j-1}) \cdot (\lambda_{1j} - \lambda_{0j}) \right\},$$

because $\lambda_{Z0} = 1$ for $Z = 0, 1$. Note that $\lambda_{1j} \geq \lambda_{0j}$ by Assumption 2 and that $\lambda_{1j}$ and $\lambda_{0j}$ equal 0 or 1. Therefore, $\lambda_{1j} - \lambda_{0j}$ equals 0 or 1, and we can write the previous expression as

$$\sum_{j=1}^{J} E[Y_j - Y_{j-1} | \lambda_{1j} - \lambda_{0j} = 1] \cdot \Pr(\lambda_{1j} - \lambda_{0j} = 1)$$

$$= \sum_{j=1}^{J} E[Y_j - Y_{j-1} | S_1 \geq j > S_0] \cdot \Pr(S_1 \geq j > S_0).$$

Similarly, for the denominator, $S = Z \cdot S_1 + (1-Z) \cdot S_0$ and, because $j$ plays the role played by $Y_j$ in the numerator,

$$E[S|Z=1] - E[S|Z=0]$$

$$= E\left\{ \sum_{j=0}^{J} j \cdot (\lambda_{1j} - \lambda_{1j+1} - \lambda_{0j} + \lambda_{0j+1}) \right\}$$

$$= E\left\{ \sum_{j=1}^{J} (\lambda_{1j} - \lambda_{0j}) \right\} = \sum_{j=1}^{J} \Pr(S_1 \geq j > S_0).$$

### Proof of Theorem 2

The denominator of the formula for $\mu_k$ is the same as the denominator of the expression for $\beta_z$. To evaluate the numerator, we can write

$$E[Y|Z=l]$$

$$= \beta_{l,l-1}(E[S|Z=l] - E[S|Z=l-1]) + E[Y|Z=l-1]$$

$$= \sum_{k=1}^{l} \beta_{k,k-1}(E[S|Z=k] - E[S|Z=k-1]) + E[Y|Z=0]$$

and

$$E\{Y \cdot (E[S|Z] - E[S])\}$$

$$= E\{E[Y|Z=l] \cdot (E[S|Z=l] - E[S])\}.$$

Using the first line to substitute for $E[Y|Z=l]$ in $E\{Y \cdot (E[S|Z] - E[S])\}$, we have

$$\sum_{l=1}^{K} \sum_{k=1}^{l} \pi_l(E[S|Z=l] - E[S])\beta_{k,k-1}(E[S|Z=k] - E[S|Z=k-1])$$

$$= \sum_{k=1}^{K} \sum_{l=k}^{K} \pi_l(E[S|Z=l] - E[S])\beta_{k,k-1}$$

$$\times (E[S|Z=k] - E[S|Z=k-1]).$$

This establishes the right side of formula for the weights, $\mu_k$. The weights are nonnegative, because the points of support of $Z$ are ordered so that $E[S|Z=k] > E[S|Z=k-1]$. To show that the weights sum to 1, note that the sum of the numerator of each $\mu_k$ is

$$\mu_k = \sum_{k=1}^{K} \sum_{l=k}^{K} \pi_l(E[S|Z=l] - E[S])(E[S|Z=k]$$

$$- E[S|Z=k-1]).$$

Reversing the order of summation as before, this equals

$$\sum_{l=1}^{K} \sum_{k=1}^{l} \pi_l(E[S|Z=l]-E[S])(E[S|Z=k]-E[S|Z=k-1]).$$

Reversing the first two steps of the proof for the numerator, this is

$$\sum_{l=0}^{K} \pi_l(E[S|Z=l]-E[S])E[S|Z=l].$$

*[Received August 1992. Revised March 1994.]*

## REFERENCES


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

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

Angrist, J., Imbens, G., and Rubin, D. (1995), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, forthcoming.

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

——— (1992), "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.

Ashenfelter, O., and Krueger, A. (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review,* 84, 1157–1173.

Becker, G. (1964), *Human Capital,* Chicago: University of Chicago Press.

Card, D., and Krueger, A. (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy,* 100, 1-40.

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

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., 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, 862–880.

Heckman, J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data,* eds. J. Heckman and B. Singer, New York: Cambridge University Press, pp. 145–245.

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," in *Sociological Methodology,* ed. Clifford C. Clogg, Washington: American Sociological Association, pp. 449–484.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 1, pp. 221–233.

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

Leamer, E. E. (1988), "Discussion," in *Sociological Methodology 1988* (Vol. 18), Washington: American Sociological Association, Chapter 14.

Newey, W. (1985), "Generalized Method-of-Moments Estimation and Testing," *Journal of Econometrics,* 29, 229–256.

——— (1990), "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica,* 58, 809–838.

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.

Powers, D. E., and Swinton, S. S. (1984), "Effects of Self-Study for Coachable Test Item Types," *Journal of Educational Psychology,* 76, 266–278.

Robins, J. M. (1989a), "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,* NCHSR, U.S. Public Health Service, eds. L. Sechrest, H. Freeman, and A. Bailey, pp. 113–159.

——— (1989b), "The Control of Confounding by Intermediate Variables," *Statistics in Medicine,* 8, 679–701.

Robins, J. M., and Greenland, S. (1992), "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology,* 3, 143–155.

Robins, J. M., and Tsiatis, 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, 2609–2631.

Rosen, S., and Willis, R. J. (1979), "Education and Self-Selection," *Journal of Political Economy,* 87, S7–S36.

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

——— (1977), "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics,* 2, 1–26.

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

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

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

Theil, H. (1958), *Economic Forecasts and Policy,* Amsterdam: North-Holland.

——— (1971), *Principles of Econometrics,* New York: John Wiley.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables are Subject to Error," *Annals of Mathematical Statistics,* 11, 284–300.

White, H. (1982), "Instrumental Variables Estimation With Independent Observations," *Econometrica,* 50, 482–499.

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