Institute for Research on Poverty
Discussion Paper no. 1077–95

Instrument Selection:
The Case of Teenage Childbearing and Women's Educational Attainment

Daniel Klepinger
Center for Public Health Research and Evaluation
Battelle Memorial Institute

Shelly Lundberg
Department of Economics
University of Washington

Robert Plotnick
Graduate School of Public Affairs and School of Social Work,
University of Washington

November 1995

This study was supported by NIH grant HD29405-01A1. We thank Mark Gritz and Dick Startz for their comments, and Valerie Cerra, Paul Higgins, and Anjanette Nelson for research assistance.

Abstract

Recent research has identified situations in which instrumental variables (IV) estimators are severely biased and has suggested diagnostic tests to identify such situations. We suggest a number of alternative techniques for choosing a set of instruments that satisfy these tests from a universe of a priori plausible candidates, and we apply them to a study of the effects of adolescent childbearing on the educational attainment of young women. We find that substantive results are sensitive to instrument choice, and make two recommendations to the practical researcher: First, it is prudent to begin with a large set of potential instruments, when possible, and pare it down through formal testing rather than to rely on a minimal instrument set justified on a priori grounds. Second, the application of more restrictive tests of instrument validity and relevance can yield results very different from those based on less restrictive tests that produce a more inclusive set of instruments, and is the preferred, conservative approach when improper instrument choice can lead to biased estimates.

Instrument Selection:
The Case of Teenage Childbearing and Women's Educational Attainment

I. INTRODUCTION

All students of econometrics learn that instrumental variables (IV) methods yield consistent parameter estimates when a regressor is correlated with the error. However, recent research that has identified situations in which IV estimators are severely biased has shaken confidence in the application of this technique. A number of diagnostics have been suggested to identify applications in which inference based on IV may be misleading. These include F-tests of instrument relevance and tests of overidentifying restrictions (OIR). Such tests, though not new, have not been routinely applied to IV estimates in the past, but are rapidly becoming a standard part of the reported statistics. An obvious question arises as these tests of instrument admissibility enter common usage: What do we do if a possible instrument set fails one or more of the tests?

In general, econometricians have provided little guidance to the practical researcher who must choose a set of valid and useful instruments from a large set of potential instruments. Instrument choice in empirical work has been based almost entirely on a priori considerations: instruments are assumed to be valid if economic theory suggests that they are unlikely to be correlated with the structural error.[1] The possibility that a set of instruments attractive on a priori grounds may fail an essential diagnostic test suggests that researchers may resort to prescreening instruments or to fishing expeditions that may compromise the integrity of the reported results.

We consider possible approaches to instrument choice in a non-artificial setting: a study of the effects of adolescent childbearing on the educational attainment of young women. The effect of early childbearing on the ability of young mothers to support themselves and their children is an issue of considerable policy importance, because an early departure from formal schooling is a likely source of economic disadvantage for teenage mothers. Our OLS estimates, using a sample of over 1700 young

[1] See Mroz 1987 for an exception.

women from the National Longitudinal Survey of Youth (NLSY), suggest that childbirth before age 18 reduces final educational attainment by 1.5 years. However, human capital theory predicts that women plan their childbearing and education decisions jointly, so OLS estimation of an educational attainment equation with teenage fertility as a regressor is likely to yield biased estimates. Early work on this topic employed a small number of instruments (sometimes one) and found that the effect of fertility on education was not significantly different from zero. For our reexamination of this question, we use a conventional data source augmented by a rich set of potential instruments describing the community and policy environment. The problem that arises in this context is how to select an instrument set from such a universe.

We suggest a number of alternative techniques for choosing a set of instruments that satisfy the standard diagnostic requirements from a universe of a priori plausible candidates. We choose mechanical procedures for eliminating potential instruments to avoid discretion in the screening process, and we try to balance the competing needs to exclude instruments that may generate inconsistent estimates and to maintain enough valid instruments to yield a powerful test of the substantive hypothesis. Conditional on the instrument set, we also implement endogeneity tests to aid the interpretation of the results.

Our first result is that a just-identified model with one excluded instrument (assumed to be exogenous) generates imprecise parameter estimates and a hypothesis test with low power. Expanding the instrument set increases our ability to explain variation in fertility, and increases the precision of the estimated effect on education, which becomes significantly negative. We also find, however, that another set of problems arises in overidentified models: estimates are likely to be sensitive to the method used for choosing a set of valid instruments. In our application, point estimates of the coefficient of interest, which corresponds to years of education lost due to teenage motherhood, range from -4.5 to -1.7 as the restrictiveness of the instrument criteria changes. This range, combined with the insignificant coefficient in the just-identified model, presents a challenge in statistical inference for those charged with developing policy concerning teenage childbearing.

The sensitivity of IV estimates to alternative, a priori reasonable methods of choosing an instrument set is a disturbing result. Since the final models in each case satisfy standard diagnostic criteria, our concern that instrument choice provides an opportunity for researchers to fish for desired results seems to be a valid one. The statistical properties of IV estimators in finite samples are not well understood and the formal consequences of alternative instrument choice algorithms have only recently become a topic of interest among econometricians.

The results reported in this paper suggest two recommendations for the practical researcher facing an instrument-choice problem: First, when possible, it is prudent to begin with a large set of potential instruments and to pare it down through formal testing, rather than to rely on a minimal instrument set justified on a priori grounds. Second, the application of very restrictive tests of instrument validity and instrument relevance can yield results very different from those based on a more inclusive set of instruments, and is the preferred, conservative approach when improper instrument choice might lead to biased estimates.

II. INSTRUMENTAL VARIABLES AND THE CHOICE OF AN INSTRUMENT SET

We can begin with a single-equation linear regression model

$$y = X\beta + u \tag{1}$$

where X is an n-by-k matrix of explanatory variables and the error u is distributed $IID(0, s^2I)$. For the OLS estimator $\tilde{\beta}$ to be consistent for β, the error must be asymptotically uncorrelated with X, although there are many situations in which we expect some elements of X to be correlated with u. These include simultaneous equations models, where some elements of X are endogenous, and errors in variables. Our application, in which y is years of schooling and the matrix X contains an indicator of adolescent fertility, provides an example in which the suspected correlation is due to a joint decision process. In such cases, the technique of instrumental variables is often applied to arrive at a consistent estimate of β.

To implement instrumental variables (in this linear model, two-stage least squares) we require a matrix of acceptable instrumental variables, *Z*. Instruments are acceptable if they are correlated with *X*, but uncorrelated with the error term in (1). If *Z* is n-by-m and includes all exogenous variables in *X*, then identification of the model requires that m $\geq$ k, or that there are at least as many instruments as there are explanatory variables in (1). Every set of valid instruments *Z* will yield a consistent estimate of β, but in finite samples estimates will differ. The question of how to choose a set of instruments arises when applying instrumental variables in cases where the universe of potential instruments is large.

The first, and most important, criterion is that the instruments must be valid, in the sense of being orthogonal to the structural error. In general, a priori arguments based formally or informally on economic theory justify the choice of particular instruments, and in the case of just-identified models, this is the best one can do. If the model is overidentified (that is, there are more instruments than regressors) we can test the validity of the instrument set. The intuition behind tests of overidentifying restrictions, as they are called, can be seen by considering the alternative (2) to the specification in (1):[2]

$$y = X\beta + Z^{*}\gamma + u \qquad \textbf{(2)}$$

where Z^* is a matrix of m-k columns of *Z* whose validity as instruments is in question. If (1) is a correctly specified model, then if (2) is estimated by IV with *Z* as the matrix of instruments, *g* should equal zero: Z^* should not explain any variation in *y* not explained by *X*. If some columns of Z^* should have been included in *X* or are correlated with *u*, then *g* will not be zero. Tests of overidentifying restrictions are tests of the joint hypothesis that (1) is correctly specified and that *Z* is a valid matrix of instruments.

A number of tests to determine the validity of potential instruments are available (Godfrey 1988; Hausman 1978; Hausman and Taylor 1981; MacKinnon 1992; Ruud 1984; White 1982). Godfrey's (1988)

[2] This discussion follows that in Davidson and MacKinnon 1993.

straightforward test of overidentifying restrictions involves regressing the residuals from an IV equation on the full set of instrumental variables *Z*. The term $n*R^2$ from this regression is distributed as χ^2 with m-k degrees of freedom and is used to test the null hypothesis that *Z* is a valid set of instruments without specifying the set Z^* .

Tests of overidentifying restrictions have a long history, but in recent years concerns about the small-sample properties of IV estimators have affected the instrument-choice problem. A number of Monte Carlo studies have identified conditions in which IV estimators are severely biased and have small-sample distributions very different from their asymptotic distributions. Two completely different situations appear to give rise to this problem: a large number of good instruments, and instruments with very little explanatory power for the endogenous regressor (low relevance).

For the former case, Davidson and MacKinnon (1993) present a set of Monte Carlo results showing that increasing the number of “good” instruments may increase the small-sample bias of an IV estimator, even though it increases asymptotic efficiency. This occurs as a result of over-fitting the data; as the number of instruments approaches the sample size, the IV estimator gets closer and closer to the OLS estimator. In large cross-section and panel studies (such as ours), the number of observations is large compared to the number of instruments, so this source of finite sample bias is unlikely to be a problem.

The case in which the instruments have little relevance for the endogenous regressors has received more attention in recent years. A number of papers have identified specific cases in which the IV estimator is severely biased. Nelson and Startz (1990a, 1990b) find that, when the correlation between a single instrument and single regressor is low, the IV estimate of the coefficient on the regressor may appear to be highly significant when the true coefficient is zero. Bound, Jaeger, and Baker (1993) show that, in general, if a set of instruments is weakly correlated with the endogenous regressor, even a small correlation between the instruments and the error can seriously bias estimates. Shea (1993) and Staiger and Stock (1993) present similar results.

Use of simple diagnostic statistics, such as the R^2 of the first-stage regression, or the F-statistic on the joint significance of all instruments in a first-stage regression, has been suggested to measure instrument relevance and as a guide to the reliability of inference from an IV estimate. However, Hall, Rudebusch, and Wilcox (1994) present evidence that the use of such diagnostics as a method of prescreening potential instruments can exaggerate finite-sample biases. This occurs because "those instruments that are identified as having high relevance for the regressors in the sample are also likely to have higher endogeneity in the sample" (p. 3). The practical implications for instrument selection of this literature on finite-sample bias are at present unclear.

The application of IV methods for equation (1) requires a set of instruments that are (a) valid or exogenous, and (b) have some relevance for the endogenous regressor. As we have seen, these properties are testable (in the case of (a), in an overidentified model only), but it is not clear how such tests should be applied in an instrument-selection problem. In most applications of IV, instrument selection is based entirely on a priori grounds and no diagnostic tests are reported. Given the biases reported for IV estimates in a variety of situations, this is likely to be an important omission, and inference based on such estimates may be misguided.

Given a set of acceptable instruments, it is possible to test another assumption implicit in the choice of IV methods: that some elements of X are correlated with the error. If this assumption is incorrect, the IV estimator is still consistent, but it is less efficient than OLS. If analysts have to choose between efficiency and consistency, most would prefer consistency and employ IV methods if they are uncertain whether X is correlated with the error. However, the cost of using IV methods may be quite high if the resulting imprecision substantially lowers the power of the test on b, or if the use of invalid instruments causes bias in the IV estimates.

Endogeneity tests can help determine whether the use of IV methods is warranted. Several such tests have been developed: Hausman (1978) and Hausman and Taylor (1981) present one of the best known.[3] The Hausman test compares a set of parameter estimates known to be unbiased with another set that is more efficient but possibly biased. Using the above example, the Hausman test compares the OLS estimates $(\tilde{\beta})$ to the IV estimates $(\hat{\beta})$. Conditional on $\hat{\beta}$ being consistent, the Hausman procedure tests whether $\tilde{\beta}$ is unbiased. If the test indicates that, at some reasonable confidence level, $\tilde{\beta}$ is unbiased, the more efficient OLS estimates may be preferred to the IV estimates. As noted by Nakamura and Walker (1994), however, inferences from specification tests can be misleading and automatic acceptance of test results may be unwise. Several issues should be considered with respect to interpreting test results for the significance of instrumental variables estimates and endogeneity tests. First, both the IV estimates and the endogeneity tests are conditional on the set of instruments chosen. If the instruments are correlated with the error term, the IV estimates may be more biased than OLS and the endogeneity tests will be biased. Second, the instruments may not be adequately correlated with the endogenous regressor, resulting in specification tests with low power. If the power of the tests is low, a false null hypothesis—that the IV estimates are equal to zero or that the endogenous regressor is exogenous—may be accepted. Finally, in the case of the endogeneity tests, we are testing the null hypothesis that the potentially endogenous regressor is exogenous. Failure to reject this null hypothesis is subject to Type II errors, which typically are not bounded in the way that Type I errors are. Thus, the probability of accepting the null hypothesis of exogeneity may be much higher than the analyst suspects.

In the next section, we propose a number of alternative procedures for instrument selection that are responsive to the possible sources of bias outlined above. First, tests of overidentifying restrictions are used to identify sets of valid instruments. These are followed by tests of instrument relevance, in response to the Bound, Jaeger, and Baker, and Nelson-Startz results. The prescreening bias that may result from selecting instruments based on relevance alone should be attenuated by this two-step procedure. We note

[3] For example, see Hausman 1978; Hausman and Taylor 1981; Ruud 1984; White 1982 and Wu 1973. For a summary of these measures, see Godfrey 1988 or MacKinnon 1992.

that prescreening tests may be applied to the set of potential instruments jointly or individually, and examine the sensitivity of resulting IV estimates to this choice.

III. TESTING, TESTING. . .

We propose several multistep approaches for selecting acceptable instruments in an empirical setting. Instruments are considered acceptable if they are both *valid* according to a test of overidentifying restrictions and *relevant* in explaining the endogenous regressor. We assume that there is a rich set of theoretically plausible potential instruments, but that the a priori arguments for their acceptability are not so compelling as to preclude testing. This is becoming a common situation in labor economics and related fields, as data on a variety of community characteristics, local economic conditions, and policy environments are appended to the individual records of cross-section and panel data sets. We assume that there are enough instruments assumed exogenous to identify the model, but we wish to add more valid instruments to generate more efficient estimates and sufficient power to test the substantive hypothesis.

The first step is to choose a set of valid instruments from the full set using a test of overidentifying restrictions (OIR). Following Godfrey (1988), we regress the residuals calculated using the IV estimates of the coefficients in (1) on all the instruments. The R^2 for this equation times the sample size is distributed as x^2 with m-k degrees of freedom. The test itself is straightforward, but its use in constructing a set of valid instruments is not. If the full set of theoretically plausible instruments fails the OIR test, there are at least three general ways we could proceed:

(1) We could exclude the instrument with the highest t-statistic from the set of potential instruments and rerun the OIR test, and repeat until a subset of the instruments passes the test. This procedure involves using the test of overidentifying restrictions as a joint test applied to the entire instrument set.

(2) We could exclude all instruments that achieved some fixed level of significance in the initial regression. This individual screening of suspicious instruments is much more restrictive and is likely to result in excluding more instruments than the joint test.

(3) We could divide the instrument set into classes based on our a priori expectations about their exogeneity (for example, put state policy variables into one class, county-level measures of medical services into another) and use the results of the initial test to update our expectations. We could then exclude an entire class of instruments if one or more members would be excluded on the basis of an OIR test.

Any of these methods will result in a set of instruments that passes a test of overidentifying restrictions, but they may result in different instrument sets, and different estimates of the coefficient of interest in finite samples. (We pursue the first two approaches in this paper, deferring the third to future work.)

The second step is to use goodness-of-fit tests to determine whether a set of valid instruments is relevant to the endogenous regressors. Simple F or likelihood-ratio tests can be used to test whether the selected instruments significantly improve model fit in the first-stage estimation. With a large number of instruments, we cannot test all possible combinations. The best testing procedure would be a mechanical one that allows systematic consideration of a large number of possible predictive models and eliminates unintended investigator bias in selecting the instruments in the final first-stage model. An approach that satisfies these criteria is to use backward-stepwise regression to select the best-fitting model that passes the goodness-of-fit test. Backward-stepwise methods are preferred to forward-stepwise methods because they tend to yield better-fitting final models when the set of potential instruments is not orthogonal. As an individual counterpart to this joint test of instrument relevance, we apply backward-stepwise regression until each identifying instrument remaining in the first-stage model achieves some set level of significance in the first-stage equation.

Once a final set of acceptable instruments is selected, an endogeneity test determines the likelihood that IV estimation is warranted on the basis of correlation between the regressor and the error. We employ Hausman's (1978) alternative test. This involves estimating the first-stage equation for the possibly endogenous regressor X_1, saving the estimated residuals $\hat{\varepsilon}$, and including them as regressors in the second-stage equation:

$$Y = X\beta + \alpha\hat{\varepsilon}. \tag{3}$$

A test of $a = 0$ tests the hypothesis that X_1 is exogenous. Hausman suggests constructing an x^2 test but an F- or t-test (if X_1 is of dimension 1) is asymptotically equivalent (Godfrey 1988; MacKinnon 1992; Vella 1993). An equivalent test for the case in which the first-stage model is a probit consists of calculating the generalized residuals from the probit and including them in the second-stage equation (Card and Vella 1994).

Following the procedures outlined above presupposes a large set of potential instruments. In a just-identified model one must rely on a priori arguments to justify exclusion restrictions. The relevance, but not the validity, of the identifying instruments can be tested. In the application presented below, a just-identified model passes a test of instrument relevance but generates estimates that are very imprecise in comparison to those produced by a more extensive set of instruments. The substantive inferences based on these contrasting results are likely to be very different, though no diagnostic criteria are violated by the former. The endogeneity test is also conditional on the instrument set chosen, and may fail to reject the exogeneity of X_1 if the instrument set cannot generate a test of sufficient power. The prudent course in IV estimation is to begin with an extensive set of instruments when possible and apply diagnostic tests, rather than to begin with a parsimonious model justified on a priori grounds.

IV. AN EMPIRICAL EXAMPLE

We use an empirical example to demonstrate the alternative approaches for selecting a valid and relevant instrument set. We show that failing to employ tests of OIR, of instrument relevance, and of endogeneity can affect substantive conclusions, and that these conclusions may be sensitive to the particular method used in selecting a valid instrument set.

The example considers the impact of teenage childbearing on women's educational attainment. This issue provides fairly typical problems researchers may encounter in estimating models with endogenous regressors. This issue is also apt because a relatively large body of research has generated estimated fertility effects that vary widely, and therefore has not resolved the debate concerning the impact of teen childbearing. A number of theoretical perspectives predict that childbearing will affect educational attainment. Since children increase the opportunity costs of remaining in school, women with children are less likely to continue in school than similar women without children. The effect may be particularly large for women less than 18 years old because they have not yet finished high school.

Early empirical work treated fertility as exogenous and found large negative impacts of a teen birth on various measures of educational attainment.[4] However, human capital theory predicts that women who intend to have children early in their life rationally choose to invest less in education because the expected return on that investment would be lower for them than for other women. Thus, human capital theory predicts that women plan their childbearing and education jointly. Regardless of whether educational intentions directly affect fertility intentions, the presence of unobservable determinants (for example, intentions regarding both fertility and education, fecundity, and other opportunity costs) of observed educational and fertility outcomes imply that fertility is likely to be correlated with the error term in the education equation. Both arguments imply that OLS estimates of an education equation containing fertility as a regressor will be biased.

Further research on the impact of teenage childbearing on educational attainment has addressed the potential endogeneity of fertility by either controlling for unobserved heterogeneity or using IV methods.

[4] Waite and Moore 1978. Also see the summary in Hofferth 1987.

This paper focuses on the IV solution.[5] Instrumental variable approaches attempted to estimate the impact of a teen birth by relying on small sets of instrumental variables, and sometimes on only one instrument: age at menarche. These studies yielded insignificant coefficients on the instrumented teen fertility variable.[6] Researchers and policy analysts tended to conclude that a teen birth has no impact on educational attainment or, more cautiously, that earlier OLS estimates were upwardly biased. Typically, these studies did not formally test the validity of the selected instruments or the endogeneity of teenage childbearing.

Sample and Variables

The data were obtained from the National Longitudinal Survey of Youth (NLSY), the Alan Guttmacher Institute (AGI), and other public sources. In 1979 the NLSY interviewed 12,686 male and female youths who were between the ages of 14 and 21 on January 1, 1979. In the version of the NLSY available for this study, re-interviews were conducted in succeeding years through 1991. The sample for this analysis includes all white women aged 14 to 20 in 1979, excluding those in the special military subsample and the oversample of economically disadvantaged whites. Adolescent fertility is defined as a birth before the respondent’s 18th birthday. Educational attainment is measured as completed years of schooling at time of interview in the year the respondent turned age 25 or, if this is missing, years of schooling at age 26. We have reported elsewhere results for other race-ethnic groups and for other measures of adolescent fertility.[7] Here we have chosen a single sample and specification to illustrate the effects of alternative instrument selection procedures.

The education equations include a large set of personal and family background characteristics as well as measures of employment opportunities. Table 1 lists the variables and their means. Personal and

[5] See Geronimus and Korenman 1992 for an example of empirical work employing heterogeneity corrections.

[6] Examples of studies that found no significant effect of fertility on education are Olsen and Farkas 1989; Ribar 1992, 1994; and Moore et al. 1993.

[7] See Klepinger et al. 1995.

family background variables include highest grade completed by mother and father, a set of variables for different living arrangements experienced as a child, number of siblings and older siblings, whether there was an adult female working for pay in the household when the respondent was age 14, whether the respondent or her parents were born outside the United States, whether the respondent was born in the South, whether the respondent lived in the South or an urban area at age 14, whether a non-English language was spoken at home when the respondent was age 14, whether her household subscribed to magazines or newspapers, whether anyone in her household had a library card, the respondent's religious affiliation (omitted category is "none"), and frequency of attendance at religious services (omitted category is "never"). We measure employment opportunities open to adolescents by the percentage of workers employed in services and in wholesale and retail trade for the state where the respondent lived at age 14. A set of county-level variables is also included, and measures aspects of the distribution of income, religious affiliation, education, and school enrollment in the county in which the respondent resided in 1979.

Instruments included in the fertility equation but excluded from the education model are shown in the lower part of Table 1 along with their means and sources. One instrument is maintained to be exogenous: age at menarche is an individual characteristic likely to affect fertility but not educational attainment. State policy variables likely to affect fertility are measured for the state in which the respondent resided at age 14. These include the maximum AFDC payment for a family of two, the presence of restrictive abortion provisions, the ages at which parental consent is no longer needed for a young woman to have an abortion or use contraception, and similar variables indicative of state policies on abortion and family planning funding and services. We measure the state-level instruments at age 14, when residential location can be regarded as

TABLE 1
Sources and Descriptive Statistics

Variables	*Mean*	*Source*
Endogenous		
Years of schooling at age 25	13.2	NLSY
Birth before age 18	.07	NLSY
Exogenous: Education and Fertility Models		
Mother's education	11.6	NLSY
Mother’s education missing	.03	
Father's education	11.6	NLSY
Father's education missing	.06	
Living arrangements at age 14		NLSY
Mother only	.08	
Mother and stepfather	.07	
Other	.05	
Both parents	.80	
Years with mother only	.66	NLSY
Years with mother and stepfather	.52	NLSY
Years in other living arrangements	.31	NLSY
Ever experienced divorce	.12	NLSY
Number of siblings	3.1	NLSY
Number of older siblings	1.8	NLSY
Number of older siblings missing	.06	
Mother worked	.52	NLSY
Foreign born	.02	NLSY
Mother foreign born	.04	NLSY
Father foreign born	.04	NLSY
Foreign language at home	.07	NLSY
Born in South	.25	NLSY
South residence at age 14	.25	NLSY
Urban residence at age 14	.75	NLSY
Magazines in home at age 14	.75	NLSY
Newspapers in home at age 14	.89	NLSY
Library card at age 14	.80	NLSY
Employment in state of residence at age 14		NLSY
Percentage in services	.18	
Percentage in wholesale/retail trade	.22	
Percentage in other	.60	
Religion		NLSY
Baptist	.16	
Catholic	.31	
Other Protestant	.30	
Jewish/Other	.13	
None	.10	
Attendance at religious services		
Never	.17	NLSY
Rare	.27	
Occasional	.19	
Often	.37	

(table continues)

TABLE 1, continued

Variables	***Mean***	***Source***
County-level Variables		
Educational spending per 1000 students	1623	CCDB
Median household income in 1979	17320	CCDB
Median gross rent in 1980	233	CCDB
Percentage of population moved into county	10.0	CCDB
Proportion of county population		CCM
Catholic	.22	
Conservative Protestant	.20	
Jewish and other	. 003	
Percentage of county population		
Education 12 or more years	67	CCDB
Education 16 or more years	15	CCDB
Percentage of families female-headed	13	CCDB
Percentage of labor force female	42	CCDB
Percentage of children in poverty families	14	CCDB
Unemployment rate in 1980	6.8	CCDB
School enrollment rate: 5–17 year olds	.78	CCDB
Proportion of 16–17 year olds in school (state)	.90	CENS
Proportion of 18–19 year olds in school (state)	.52	CENS
Excluded Instruments		
Individual		
Age at menarche	12.9	NLSY
State level		
Maximum AFDC payment to 2 person family	$211	HEW1
Restrictive abortion provisions	.07	HEW2
Restrictive laws on the sale/advertisement of contraception	.41	HEW2
Restrictions on Medicaid funding of abortion	.20	HEW2
Maximum percent of state median income for eligibility under Title XX family planning services	.54	HEW2
No maximum	.02	
Age of consent for abortion	16.4	HEW2
No age of consent	.64	
Age of consent for contraception	16.5	HEW2
No age of consent	.69	
County level		
Abortion rate per 1000 women	21.4	AGI
Abortion provider providing more than 400 abortions	.49	AGI
Presence of abortion provider	.70	AGI
Proportion of women 15–19 using family planning services	.13	AGI
Proportion of family planning patients aged 15–19	.35	AGI
Family planning clinics per 1000 women aged 15–19	.43	AGI
Number of patients per family planning clinic	1363	AGI
Hospital expenditures per 1000 population	48	CCDB
Number of doctors per 1,000,000 population	1590	CCDB
Number of nurses per capita	.005	CCDB

(table continues)

TABLE 1, continued

Variables	***Mean***	***Source***
County level fertility rates and sex ratio		
Marital fertility rate women aged 15–19	370	AGI
Nonmarital fertility rate women aged 15–19	16	AGI
Ratio of men 15–19 to women 15–19	94.6	AGI
Number of observations	1728	

Notes:

NLSY: Data were obtained from the National Longitudinal Survey—Youth Cohort.

AGI: Data were obtained from the Alan Guttmacher Institute.

HEW1: Data were obtained from the U.S. Department of Health, Education, and Welfare.

HEW2: Data were prepared for the U.S. Department of Health, Education, and Welfare by the Alan Guttmacher Institute.

CCDB: Data were obtained from the City-County Data Book.

CCM: Data were obtained from Quinn et al.1982.

CENS: Data were obtained from the 1980 Census of the United States.

exogenous. County-level indicators of the availability of abortion and family planning services are likely to be good measures of the costs of abortion and contraception, but they may also reflect local demand for such services and may be correlated with unobserved local norms and attitudes that also affect educational decisions. We measure these instruments for the county in which the respondent was living at the time of interview in 1979 (or in 1980 if data are not available for 1979). We would prefer to measure these variables at uniform early age (as we did for the state-level ones), but county of residence is not available prior to 1979. The county-level instruments are the abortion rate, whether there is an abortion clinic performing more than 400 abortions, the proportion of women aged 15 to 19 using family planning services, hospital expenditures, and similar variables listed at the end of Table 1. Finally, measures of the social context within which fertility decisions are made include marital and nonmarital fertility rates for women aged 15 to 19 and the ratio of young men to young women in the county.

Econometric Procedures and Findings

We consider three estimators: ordinary least squares (OLS), in which observed fertility by age 18 is assumed exogenous; the standard two-stage least squares (2SLS) procedure that employs an OLS first-stage for birth by age 18; and a two-stage probit-regression procedure (2SPR) that uses the predicted probability from a first-stage probit as the excluded instrument for the linear second stage.[8]

Table 2 displays the key parameter estimates for these estimators and results from the statistical tests discussed earlier.[9] Row 1 displays the OLS point estimate for the effect of having a birth prior to age 18 on highest grade completed at age 25. The second and third rows present corresponding 2SLS and 2SPR

[8] The 2SLS estimator is consistent when the stochastic regressor is dichotomous (see Heckman 1978 for a discussion). The 2SPR model is identical to a model of treatment effects, where the predicted probability of a teen birth is the only excluded instrument in the model.

[9] In addition to the observed or instrumented measure of fertility, all models include the large set of personal, family background, and economic control variables reported in Table 1. The second-stage regressions have R^2 of about .35. Complete regression results are available upon request.

TABLE 2

Effect of a Birth Before Age 18 on Educational Attainment: Instrument Choice with Tests of Overidentifying Restrictions and Instrument Relevance

	Method for Choosing Instruments					
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Menarche	*Full set of*	*Joint*	*Joint*	*Joint Relevance*	*Individual*
	only instrument	*instruments*	*Relevance*	*OIR*	*& OIR*	*Relevance &*
OIR						
Coefficients on Fertility Dummy [a]						
OLS	-1.56 (.17)**					
Two-Stage Least Sq. (2SLS)	-2.19 (3.01)	-4.24 (1.34)**	-3.59 (1.29)**	-4.50 (1.38)**	-4.45 (1.38)**	-2.66 (1.58)*
Two-Stage Probit (2SPR)	-2.04 (1.37)	-2.47 (.88)**	-2.19 (.89)**	-1.76 (.92)*	-1.99 (.93)**	-1.68 (1.02)*
Overidentifying Restrictions (OIR) (χ^2 Test)[b]						
2SLS		41.82 (23)**	26.70 (17)*	21.77 (21)	18.66 (16)	.69 (5)
2SPR		49.59 (23)**	30.07 (17)**	26.96 (21)	12.79 (13)	1.21 (5)
Instrument Relevance (F-Test) [b]	4.89 (1)**	1.26 (24)	1.65 (18)**	1.32 (22)	1.71 (17)**	3.21 (6)**
Endogeneity of Fertility (t-test)						
2SLS	.21	2.18**	1.56	2.34**	2.29**	.71
2SPR	.35	1.93**	.71	.57	1.33	.92

* = Significant at the 10 percent level.
** = Significant at the 5 percent level.

[a] Standard errors in parentheses.
[b] Degrees of freedom in parentheses.

estimates. Rows 4 and 5 present the tests of overidentifying restrictions—the x^2 values for the regressions of the 2SLS and 2SPR error terms on the set of instrumental variables (instruments and controls). Row 6 displays the F-statistic for testing the improvement in model fit when the selected instruments are added to an OLS model of childbearing prior to age 18 which includes the extensive set of control variables. The last two rows display the t-statistics for the tests of whether birth prior to age 18 is exogenous.[10]

For comparisons with prior empirical work on the impact of a teen birth, the results in column 1 use age at menarche as the only instrument. As shown in the first row, the OLS estimate is statistically significant (t = 9.2) and large. A birth prior to age 18 is associated with more than 1.5 years less education by age 25. The 2SLS and 2SPR estimates are not statistically significant, although the point estimates are somewhat larger than in row 1. The F-test indicates that including age at menarche in the first-stage regression significantly improves the fit. Age at menarche passes the diagnostic test for instrument relevance suggested in Bound, Jaeger, and Baker. The last two rows indicate that we cannot reject the null hypothesis that having a child by age 18 is exogenous in the just-identified model.[11]

What can the analyst conclude from column 1? If we consider only the IV estimators, we might conclude that a teen birth does not affect educational attainment. However, the endogeneity tests imply a small probability that fertility is endogenous, and so it is not clear that we should base inference on the IV estimates, rather than the OLS results. The OLS estimates show that early fertility has a significant, large negative impact on educational attainment.

How can we reconcile these contradictory test results? The answer lies with the power of the tests. Although the statistical significance of age at menarche in predicting fertility suggests that it is an acceptable instrument, the IV estimators have standard errors about 8, that is, 17 times larger than the one

[10] Godfrey (1988), MacKinnon (1992), and Vella (1993) point out that in this form of the Hausman test, the F-test (or t-test, with one potentially endogenous regressor) is equivalent to the chi-square test suggested by Hausman and is easier to calculate since the necessary information for calculation is part of the standard output.

[11] Since age at menarche is the only instrument, the model is exactly identified. Consequently, tests of overidentifying restrictions cannot be applied because the test statistic will have zero degrees of freedom.

from the OLS model. While it is not surprising that IV procedures produce larger standard errors than OLS, the difference is quite large. Moreover, the standard error for the more efficient 2SPR is 1.37, implying that the t-test for birth before age 18 is not powerful enough to detect effects at the .05 level of less than 2.68 years (2.68/1.37 = 1.96). Similarly, the endogeneity test based on the just-identified model is not powerful enough to detect effects at less than 2.04 years deviation from OLS. Evidently, using age at menarche as the sole instrument does not generate statistical tests with sufficient power to assess the impact of fertility on educational attainment or the endogeneity of fertility.

An obvious solution to this problem is to obtain more instruments. Column 2 displays results when we use the full set of 24 excluded instruments from Table 1. The point estimates in rows 2 and 3 are significant. They are larger than the OLS estimate and larger than those obtained when age at menarche is the only instrument. If the analyst considers only the results with the full instrument set, the conclusion would be that having an early birth reduces a woman's educational attainment by between 2.5 and 4.2 years—the latter a very large estimate. Both the 2SLS and 2SPR models fail the test for overidentifying restrictions, suggesting that this instrument set may be overinclusive. The F-test in row 6 indicates that the full set of instruments does not significantly improve the predictive ability of the first-stage regression, so the small sample bias discussed by Bound, Jaeger, and Baker, and Nelson and Startz may be a problem. Finally, the endogeneity tests in rows 7 and 8 indicate that we can reject the null hypothesis of exogenous fertility for both models.

To examine the separate effects of failing the tests of overidentifying restrictions and instrument relevance, we reestimate the IV models using different criteria for selecting the final instruments. Column 3 displays results when we use only the test of instrument relevance to select acceptable instruments. Starting with the full set of 24 potential instruments, we use backward-stepwise regression to delete variables until we obtain a first-stage model that passes the joint goodness-of-fit test. Removing 6 instruments achieves this goal. The resulting point estimates in rows 2 and 3 are somewhat smaller than those in column 2, but again are statistically significant. However, the tests for overidentifying restrictions in rows 5 and 6 reject the hypothesis that the 18 instruments passing the goodness-of-fit test are uncorrelated with the error term. Hence, this subset of 18 instruments may still yield biased estimates.

Column 4 displays results based on an instrument set that passes the overidentifying restrictions test at a .10 level. To identify this set, we start with all 24 instruments and exclude the one with the highest t-value in the regression of the second-stage residuals on the set of selected instruments and control variables. We then reestimate the IV model without this instrument. This process is repeated until we obtain a model that passes the test of overidentifying restrictions. In this case, the final set includes 22 instruments. The difference between the resulting fertility coefficients and those in column 3 is statistically significant: the 2SLS coefficient increases in absolute value, while the 2SPR coefficient falls. This set of instruments fails the instrument relevance test.

Column 5 shows results for models based on a set of instruments that passes *both* the instrument relevance test *and* the test of overidentifying restrictions. To select this set, we first regress the second-stage residuals on the full set of instruments to test the overidentifying restrictions. If the model fails the test at a .10 significance level, we exclude the instrument with the highest t-test in the regression and retest the new subset. This procedure is followed until we find a set of instruments that pass the test. Beginning with this subset of instruments, we estimate a backward-stepwise regression model for fertility and delete variables until we obtain a model that passes the joint relevance test. Finally, we apply the test of overidentifying restrictions to this new subset to confirm that we have a set of instruments that passes both criteria of acceptability.

This process yields a 2SLS model based on 17 instruments, a 2SPR model with 14 instruments, and estimates that are not significantly different from those obtained using the full set of instruments. The exogeneity of fertility is rejected in the 2SLS model. Based on the 2SLS model in column 5, the analyst would conclude that fertility is endogenous and that having a birth before age 18 is associated with 4.5 years less completed education. The estimated substantive effect is rather high, but a reasonably prudent investigator is likely to stop with these results, based on instruments that pass the standard diagnostic tests.

Columns 3, 4, and 5 are based on the application of *F* and x^2 tests to groups of instruments. Using a joint test to determine the acceptability of a set of instruments may fail to exclude individual instruments that are invalid or of low relevance. The joint test for overidentifying restrictions tests whether a set of instruments is correlated with the error term. A potential instrument that is highly correlated with the error term might be included among a group of variables that jointly pass the test if the other variables have very low correlations with the error term. Similarly, a potential instrument with little predictive power may escape exclusion under a joint relevance test if other instruments in the group have considerable predictive power. In an empirical setting, we do not know whether the inclusion of such instruments is likely to result in a biased IV estimator.

An alternative approach for identifying acceptable instruments is to test each one individually. To explore whether individual tests and joint tests generate different results, we estimate IV models that include only instruments that pass the acceptability criteria individually. For the results reported in column 6, we excluded a potential instrument if its coefficient in the overidentifying restrictions regression was significantly different from zero at a .10 significance level or if its coefficient in the regression predicting fertility was *not* significantly different from zero at the .10 significance level. These individual tests are more conservative than the joint tests and consistently yield fewer acceptable instruments.[12]

Only 6 instruments pass muster in column 6. The IV point estimates suggest that having a birth prior to age 18 is associated with acquiring 1.7 to 2.7 fewer years of schooling by age 25. The point estimates in column 6 are significantly different from those produced by the full instrument set (column 2) and, for the linear model, by the joint tests in column 5. Further review of column 6 suggests a surprising, perhaps awkward, conclusion. Rows 7 and 8 do not reject the hypothesis that fertility is exogenous with respect to educational attainment, which, ironically, implies that the original OLS estimate is unbiased and the IV results are not needed. Although the endogeneity test for the 2SPR is capable of detecting

[12] We estimated an alternative model in which the less restrictive significance level of .20 was applied to the instrument relevance test. This resulted in an identical instrument set in this sample.

deviations from the OLS estimate in excess of 1.57 years at the .05 level, the 2SPR point estimate is nearly identical to the OLS estimate.

This stringent, formalized procedure for selecting the final set of instruments should give one confidence in the unbiasedness of the resulting IV estimates. The sensitivity of the 2SLS coefficient to a choice among instrument sets that pass standard diagnostic tests is somewhat disturbing. Since the individual tests are more conservative in excluding potentially unacceptable instruments than the joint tests, we consider the estimates in column 6 less likely to be biased than those in column 5, though they are somewhat less precisely estimated.

The results in column 6 leave the analyst with a sizable range of estimates for the effect of a birth prior to age 18 on years of education by age 25, but one that is significantly lower than that implied by the full instrument–set IV estimates in column 2. Rejection of the endogeneity of fertility in the education equation, however, brings our attention back to the precisely estimated OLS coefficient. If early childbearing is an outcome independent of the unobserved determinants of a young woman’s educational objectives, then a reasonable estimate of the penalty it exacts on educational attainment is 1.6 years, with a standard error of only .17 years.

Though the findings in Table 2 are contingent on a particular substantive issue and the data used to examine it, several lessons emerge from looking across the six columns. Choosing only enough instruments to identify the model exactly may not yield statistical tests with sufficient power to assess the substantive effect of the endogenous regressor or to test whether, in fact, it is endogenous (column 1). Amassing a large number of potential, theoretically sensible instruments may well lead to biased results if one does not test the instruments’ acceptability (column 2). Column 5 and the intermediate analyses in columns 3 and 4 seem to demonstrate that conducting tests of acceptability can give one confidence that the estimates are unbiased, while confirming the theoretical view that the regressor being instrumented is, indeed, endogenous. However, applying more restrictive tests for acceptability can result in substantively different estimates (column 6). Finally, our results support the argument made by Nakamura and Walker

(1994) that specification tests should be interpreted cautiously and that other relevant information should be considered. In particular, our results show that it is important to consider both the power of the specification tests and the quantitative size of the parameter differences generated by different estimators.

V. CONCLUSION

Models with endogenous regressors are commonplace in labor economics and other applied fields, and instrumental variables has been a popular technique for consistently estimating the parameters of interest. However, recent research has identified a number of situations in which IV estimators are severely biased. Diagnostics that have been suggested to identify applications in which inference based on IV may be misleading include F-tests of instrument relevance and tests of overidentifying restrictions. In this paper, we suggest alternative techniques for choosing a set of instruments that satisfy the standard diagnostic requirements from a universe of a priori plausible candidates. The results are sensitive to our choice among apparently reasonable selection procedures.

We apply these techniques to a question of considerable policy relevance: the effects of adolescent childbearing on the educational attainment of young women. Early work on this topic employed a small number of instruments (sometimes one) and found that the effect of fertility on education was not significantly different from zero. For our reexamination of this question, we use a conventional data source augmented by a rich set of potential instruments describing community characteristics and the policy environment. We choose mechanical procedures for eliminating potential instruments to avoid discretion in the screening process, and try to balance the competing interests in excluding instruments that may generate inconsistent estimates and in maintaining enough valid instruments to yield a powerful test of the substantive hypothesis.

Our first result is that restricting our instrument set to one, assumed exogenous, excluded instrument generates imprecise parameter estimates and a hypothesis test with low power. Expanding the instrument set increases our ability to explain variation in fertility, and increases the precision of the estimated effect on education, which is now significantly negative. Our results suggest that researchers should, when possible, devote resources to expanding the set of potential instruments beyond what is minimally required to identify the model.

We also find that eliminating potential instruments that explain little of the variation in fertility or that appear to be correlated with the error in the education equation may cause a significant decline in the absolute magnitude of the fertility coefficient. Our most restrictive technique, which excludes potential instruments that fail diagnostic tests individually, yields estimated coefficients that are reasonably precise, of consistent magnitude, and at the low end of the range of point estimates. In fact, these estimates are not significantly different from the OLS estimate, and the hypothesis that fertility is exogenous cannot be rejected in a model with a highly restrictive set of instruments.

Researchers applying IV methods in overidentified models need to choose a method for selecting instruments that avoids the biases known to be associated with invalid instruments and with low instrument relevance. Our results demonstrate the importance of considering the power of specification tests in interpreting results and selecting a model. Our results also suggest that applying more restrictive criteria in the instrument-choice process may materially affect the resulting estimates and is the preferred, conservative approach.

References

Bound, John, David Jaeger, and Regina Baker. 1993. "The Cure Can Be Worse Than the Disease: A Cautionary Tale Regarding Instrumental Variables." Unpublished manuscript, University of Michigan.

Card, David, and Francis Vella. 1995. "Testing the Validity of Instruments in Models with Endogenous Treatments." Unpublished manuscript, Princeton University.

Davidson, R., and MacKinnon, James G. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Geronimus, A., and Korenman, S. 1992. "The Socioeconomic Consequences of Teen Childbearing Reconsidered." *Quarterly Journal of Economics* 107(4):1187–1214.

Godfrey, L. G. 1988. *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press.

Hall, A. R., G. D. Rudebusch, and D. W. Wilcox. 1994. "Judging Instrument Relevance in Instrumental Variables Estimation." Unpublished paper, Federal Reserve Board.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46(6):1251–71.

Hausman, Jerry A., and William E. Taylor. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica* 49(6):1377–98.

Hofferth, S. 1987. "Social and Economic Consequences of Teenage Childbearing." In *Risking the Future*, vol. 2, edited by S. Hofferth and C. Hayes, pp. 123–44. Washington, D.C.: National Academy Press.

Klepinger, Daniel, Shelly Lundberg, and Robert Plotnick. 1995. "Adolescent Fertility and the Educational Attainment of Young Women." *Family Planning Perspectives* 27 (1):23–28.

MacKinnon, James G. 1992. "Model Specification Tests and Artificial Regressions." *Journal of Economic Literature* 30(1):102–46.

Moore, K., D. Myers, D. Morrison, C. Nord, B. Brown, and B. Edmonston. 1993. "Age at First Childbirth and Poverty." *Journal of Research on Adolescence* 3(4):393–422.

Mroz, Thomas A. 1987. “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions.” *Econometrica* 55(4):765–99.

Nakamura, Alice, and James R. Walker. 1994. “Model Evaluation and Choice.” *Journal of Human Resources* 29(2):223–47.

Nelson, C. E., and R. Startz. 1990a. “Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator.” *Econometrica* 58 (July):967–76.

Nelson, C. E., and R. Startz. 1990b. “The Distribution of the Instrumental Variable Estimator and Its T–Ratio When the Instrument Is a Poor One.” *Journal of Business* 63 (part 2):S125–S40.

Olsen, R., and G. Farkas. 1989. “Endogenous Covariates in Duration Models and the Effect of Adolescent Childbirth on Schooling.” *Journal of Human Resources* 24(1):39–53.

Quinn, B., H. Anderson, M. Bradley, P. Goetting, and P. Shriver, eds. 1982. *Church and Church Membership in the United States, 1980: An Enumeration by Region, State, and County, Based on Data Reported by 111 Church Bodies*. Atlanta: Glenmary Research Center.

Ribar, David C. 1992. “A Multinomial Logit Analysis of Teenage Fertility and High School Completion.” *Economics of Education Review* 12(2):153–64.

Ribar, David C. 1994. “Teenage Fertility and High School Completion.” *Review of Economics and Statistics* (76):413–24.

Ruud, Paul A. 1984. “Tests of Specification in Econometrics.” *Econometric Reviews* 3(2):211–42.

Shea, J. 1993. “Instrumental Relevance in Linear Models: A Simple Measure.” SSRI Working Paper No. 9312, University of Wisconsin–Madison.

Staiger, D., and J. H. Stock. 1993. “Asymptotics for Instrumental Variables Regressions with Weakly Correlated Instruments.” Unpublished paper, Harvard University.

Vella, Francis. 1993. “A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors.” *International Economic Review* 34(2):441–57.

Waite, L. and K. Moore. 1978. “The Impact of an Early First Birth on Young Women’s Educational Attainment.” *Social Forces* 56 (March):845–65.

White, Halbert. 1982. “Maximum Likelihood Estimation of Misspecified Models.” *Econometrica* 50(1):1–26.

Wu, De-Min. 1973. “Alternative Tests of Independence between Stochastic Regressors and Disturbances.” *Econometrica* 41(4):733–50.