The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation[1]

Michael P. Murray
Bates College
October, 2006

ABSTRACT

Instrumental variables estimation can, in principle, avoid biases that ordinary least squares estimation suffers when explanatory variables are correlated with the disturbances. Finding appropriate instruments is a challenge. This paper uses nine recently published empirical papers to illustrate exemplary practices in IV estimation. Nine strategies for avoiding bad instruments (those correlated with the disturbances), as well as recently developed best practices for coping with weak instruments (those little correlated with the troublesome explanatory variable), are summarized and illustrated. The ugly interpretive perils posed by heterogeneity in agents' behavioral responses to a troublesome explanator are also described and illustrated. All procedures recommended in the paper can be implemented using existing commands (some of them quite newly constructed) for one or more standard econometric packages.

[1] Daron Acemoglu, Josh Angrist, Manuel Arellano, Bill Becker, Denise DiPasquale, Jerry Hausman, Jim Heckman, Simon Johnson, Peter Kennedy, Jeff Kling, Marcelo Moreira, Jack Porter, Carl Schwinn, and Jim Stock have provided helpful comments on the paper as a whole. Janet Currie, Carolyn Hoxby, Larry Katz, and Motohiro Yogo have helped me accurately portray their work. I am grateful to Vaibhav Bajpai for able research assistance.

Archimedes said "Give me a place to stand and a long enough lever and I can move the world." Finding the place to stand was, of course, the hard part. So it is with economists' popular lever, the instrumental variable (IV) estimator. Where do we find a variable that is correlated with a specific troublesome explanator and not with the disturbance term? How do we know when we have succeeded or failed in this quest? This article reports how a number of careful economists have tried to avoid bad instruments (those correlated with the disturbances), weak instruments (those too little correlated with the troublesome explanator), and ugly instruments (those that yield results uninformative about what we are interested in). Collectively, their efforts provide the rest of us with a guide to instrumental variables estimation.[2]

This article discusses the findings and methods of nine published studies which I chose for their important topics and because they offer illustrations of exemplary econometric practices. The substantive economic lessons of these studies highlight how much leverage good instruments can give us in uncovering how the economic world works. Section 1 of this paper reviews the instruments these economists devised to obtain consistent estimates of economic parameters of interest.

Textbooks have long underscored that valid instruments must be uncorrelated with the disturbances. Concerns about the adverse consequences of using weak instruments have become prominent more recently, as have concerns about the proper interpretation of instrumental variables estimates. Sections 2, 3, and 4 of this paper describe how bad instruments, weak instruments, and ugly instruments can make IV estimation a poor empirical choice, and illustrate how such perils might be avoided.

To anticipate what is to come, current good practice in using an instrumental variable has seven steps (all of which can be executed with one or more existing econometrics packages):

i. Check the significance and estimated signs of the instruments in the troublesome variable's reduced form and in the dependent variable's reduced form for consistency with the instrument's rationale.
ii. Avoid bad instruments. Build a case for the validity of the instrument, using the nine strategies described and illustrated below.
iii. Test for a weak instrument using Stock-Yogo critical values (Stock and Yogo (2004)). If weakness is not rejected, proceed to (iv). If weakness is rejected, either turn to 2SLS or, preferably, proceed to (iv).
iv. Conduct hypothesis tests using the conditional likelihood ratio (CLR) test (Moreira (2003)) or its robust variants.[3]
v. Build confidence intervals based on the CLR test or its robust variants.

[2] For an introductory modern treatment of IV estimation, see Murray (2006), Chapter 13, or Stock and Watson (2003), Chapter 10.

[3] When instruments are weak and there are multiple troublesome explanators, the CLR test is not appropriate when testing a hypothesis about just a subset of those coefficients. In this case, all known tests suffer size distortions, and which is the best procedure is unknown. A similar caveat applies to building confidence intervals.

vi. Obtain point estimates using Fuller's estimators with a = 1 or 4 (Fuller (1977)) rather than using 2SLS.
vii. Interpret the IV results with care. Know whose behavior your instrument identifies.

Each of these steps is described and illustrated below. They serve well as long as the estimated equation's degree of over-identification is low relative to the sample size.

1. Choosing Promising Instruments

Ordinary Least Squares (OLS) is inconsistent when one or more explanators in a regression are contemporaneously correlated with the regression's disturbances. A regression's explanators can be correlated with the disturbances because of an omitted explanator, a mis-measured explanator, an endogenous explanator, or a lagged dependent variable among the explanators. I call all such explanators "troublesome". IV estimation can consistently estimate regression parameters despite troublesome explanators.

In the past fifteen years or so, economists have exerted considerable attention to finding suitable instrumental variables for a wide variety of analyses. The nine IV analyses I summarize here tackle substantive economic questions with thoughtfully conceived instruments. My presentation paints the analyses in broad brush, but each of these papers offers a nuanced discussion of the channels through which their instruments are correlated with troublesome variables. In so short a space, it is impossible to assess all the strengths and weaknesses of these papers. Instead, my goal is to select from them illustrations of excellent econometric practice. I urge readers planning on doing IV estimation to examine at least a couple of these papers in their original, detailed form.

Most often, regressions requiring IV estimation have a single troublesome explanator, plus several non-troublesome explanators:

$$Y_{1i} = \beta_0 + \beta_1 Y_{2i} + X_i \beta_2 + \varepsilon_i , \qquad (1)$$

in which Y_{1i} is the dependent variable of interest, Y_{2i} is the troublesome explanator, and X_i is a vector of non-troublesome explanators. Instrumental variables estimation is made possible by a vector Z containing l variables that are purportedly (i) uncorrelated with the ε_i, (ii) correlated with Y_{2i}, and (iii) not explanators in equation (1). The elements of Z are called instrumental variables. IV estimators rely on the elements of Z to consistently estimate the coefficients of equation (1). When there is more than one instrument, there is an infinite number of potential IV estimators, each using the instruments differently.

The most frequently used IV estimator is two-stage least squares (2SLS). With one troublesome explanator, 2SLS combines the elements of Z, along with the elements of X, in Y_{2i}'s fitted value, $\hat{Y}_{2i}$, from a reduced form regression

$$Y_{2i} = \alpha_0 + Z_i \alpha_1 + X_i \alpha_2 + \mu_i . \qquad (2)$$

When there are additional troublesome variables, 2SLS combines the elements of *Z*, again along with the elements of *X*, in corresponding reduced-form fitted values, using the same reduced-form explanators as in equation (2).

When the number of instruments, *l*, equals the number of troublesome variables, *q*, we say the equation of interest is exactly identified. When $l > q$, we say the equation is over-identified.[4] The requirement that the elements of *Z* are not explanators in the equation echoes the classic simultaneous equation "order condition" for identification: to be identified, an equation must exclude at least one exogenous variable for each endogenous explanator it contains - the excluded exogenous variables are then available for inclusion in *Z*.

For a bivariate regression in which the one explanator is troublesome, the OLS and 2SLS estimators are $\sum y_{2i}y_{1i} \Big/ \sum y_{2i}^2$ and $\sum \hat{y}_{2i}y_{1i} \Big/ \sum \hat{y}_{2i}^2$, where lower case letters denote deviations from a variable's mean. The estimates from the formula for 2SLS equal estimates from the linear IV estimator $\sum \hat{y}_{2i}y_{1i} \Big/ \sum \hat{y}_{2i}y_{2i}$, in which the fitted value of Y_2 serves as Y_2*'s* instrument. The OLS and linear IV formulae highlight that OLS uses all of the troublesome variable's variation ($\sum y_{2i}^2$), while IV estimation only uses variation in the troublesome explanator that is shared in common with the instrument ($\sum \hat{y}_{2i}y_{2i}$). When OLS is consistent, it is also more efficient than IV estimation

Institutions and Growth

Daron Acemoglu, Simon Johnson, and James Robinson (AJR) ask whether institutions are important for economic growth (AJR (2001)). They report affirmative results for one particular feature of governments: the protection of private property from expropriation, as measured by the "risk of expropriation" index from the publication *Political Risk Services*. They note that other measures yield similar results.[5] AJR focus on countries that were colonized by Europeans. These countries had similar levels of per capita income 400 years ago, but vary widely in their incomes today. Does the risk of expropriation partially account for the differences in these countries' incomes?

Could an OLS regression of income on the risk of expropriation convince us that better (less extractive) institutions raise incomes? Unfortunately, no. A positive OLS-estimated coefficient on the risk of expropriation could as easily result from richer countries being more able and more inclined to "purchase" with their wealth non-extractive institutions, as from a causal effect of such institutions on income. To isolate

[4] Strictly speaking, having at least as many instruments as troublesome variables is only a necessary condition for identification. In most applications, the condition proves sufficient. However, when there are multiple troublesome variables, some additional attention should be given to ensuring identification.

[5] Edward Glaeser, Rafael La Porta, Florencio Lopez-de-Silanes, and Andre Schleifer argue that the risk of expropriation is a poor proxy for a society's institutions (Glaeser et al. 2004). AJR (2005) offers other proxies. I don't assess this debate.

the causal effect of the risk of property expropriation on income, AJR turn instead to IV estimation.

AJR argue that for most countries, today's institutions are linked to institutions of the past. They further argue that the less intensively a colonized country was settled by Europeans, the more extractive the political institutions that the colonizers imposed on the population. (Even mildly extractive polices spurred a revolution in the relatively intensively settled North American colonies.) Finally, AJR argue that proportion of Europeans in early colonial period populations depended, in part, on the expected mortality rate of European settlers in the colonized country. In consequence, the authors expect (and find) that positive traits of today's political institutions are negatively correlated with the death rates for Europeans in the country at the time the country was first colonized. Military and church records provide such death rates for European soldiers, sailors, and bishops, and the log of this mortality rate is the instrument that AJR exploit.[6,7]

Using 2SLS, AJR find large effects of the risk of expropriation on income. For example, Nigeria's per capita income is currently 9% of that in Chile. AJR estimate that if Nigeria had Chile's risk of expropriation, its income would be 70% of Chile's - a dramatic effect. AJR argue that the risk of expropriation serves as a proxy for a county's institutions, and interpret their result as indicating that less extractive institutions facilitate growth.

Incarceration and Crime

Steven Levitt asks whether the incarceration of criminals reduces crime rates (Levitt (1996)). Intuition suggests it does: locked up criminals can't commit crimes and other criminals might be deterred from crime by greater risks of incarceration. But is there really such an effect, and, if so, how large is it?

Could an OLS regression of changes in crime rates on changes in incarceration rates convince us of the magnitude of the causal effect of incarceration rates on crime rates? Unfortunately, no. Changes in current incarceration rates are probably influenced by past changes in crime rates, and changes in crime rates might well be serially correlated. OLS estimates of incarceration's effects on crime would be biased upward (toward zero) by such linkages; consequently, the largely negative OLS estimates in the literature might be too small in magnitude.

---

[6] In a related paper, AJR provide another instrument for ex-colonies' current institutions: the log of the country's population density in 1500 (AJR 2002).

[7] An attraction of this instrument, in AJR's view, is that the primary causes of a high death rate among early European settlers were malaria and yellow fever, which did not much affect the local adults, because the local adults had usually developed partial immunity during childhood. Consequently, AJR argue, the death rates among early European settlers are not a proxy for early health conditions for the non-European population.

Levitt turns to IV estimation to overcome the potential biases in OLS estimates of incarceration's effect on crime. The instruments he uses are all rooted in prison-overcrowding lawsuits that took place in a dozen states across a span of thirty years. These dozen states were sometimes involved in such suits, and sometimes not. Other states were never involved in such suits.[8] When such suits are filed, states defensively work to reduce incarceration rates. When such suits are won, there are further declines in prison populations. Levitt expected (and found) that incarceration rates and overcrowding litigation are negatively correlated. Levitt's instruments are based upon the stages of prison over-crowding lawsuits, from filing through judgment.

Using 2SLS, Levitt estimates that the effects of incarceration are 2-3 times larger in magnitude than indicated by previous OLS estimates. He estimates that the marginal benefit from incarcerating one prisoner for an additional year is $50,000. Published estimates of the costs of incarceration indicate that year costs the state about $30,000. Based on these estimates, Levitt judges that, "the current level of imprisonment is roughly efficient, though there may be some benefit from lengthening the time served by the current prisoner population." (Levitt, 1996, p. 324)

Public Housing and Kids

Janet Currie and Aaron Yelowitz ask whether living in public housing is good for kids (Currie and Yelowitz (2000)). Does "living in the projects" mean a child lives in a less crowded apartment in a less densely populated building and performs better in school than he or she otherwise would? The public perception of public housing is certainly at odds with such a claim. The general public would probably predict just the opposite, that public housing is *bad* for kids – that it puts them in more crowded apartments in more densely populated buildings and leads to worse academic outcomes for them.

OLS regressions of apartment quality, neighborhood quality, and kids' school performance on the traits of poor households in and out of public housing lends partial support to the common view of public housing. OLS estimates imply that public housing residents live in less desirable apartments and neighborhoods than do households with otherwise similar measured traits. And while their children do not repeat grades in school more often than children from otherwise similar households, kids from public housing do change schools more often than other children. However, the OLS estimates of the effect of being in public housing on these three outcomes may well be biased by unobserved heterogeneity; public housing tenants might be people with unmeasured traits that contribute to poor housing outcomes and poor academic performance, which would bias the OLS estimates of public housing's effects toward negative outcomes.

To overcome the potential biases of OLS, Currie and Yelowitz propose an instrumental variable for the dummy variable that indicates whether a child lives in

[8] Levitt restricts his attention to lawsuits against entire state prison systems. He ignores lawsuits against specific prisons. Some states whose state-wide prison systems were never challenged did experience lawsuits against individual prisons. Levitt ignores these latter suits because there is no evidence they affected state-wide prison populations.

public housing: the sex composition of the children in a household. (Angrist and Evans use such a dummy as an instrument in a study of fertility (Angrist and Evans 1998).) The size of the apartment a family is given in public housing depends on family size and the sex composition of the children. In particular, parents of a boy and a girl get a three bedroom apartment, while parents of two boys or two girls receive a two bedroom apartment. Because federal regulations set rents in public housing at twenty-five percent of the resident family's income, without regard to the traits of the occupied apartment, public housing offers parents of a boy and a girl a particularly attractive deal, relative to the deal offered parents of two boys or two girls. As a consequence, households with one boy and one girl are twenty-four percent more likely to be in public housing than are households with two boys or two girls. That is, the sex composition of households is correlated with their public housing participation. Currie and Yelowitz restrict their analysis to households with two children and use as their instrument a dummy variable that distinguishes mixed sex-composition children households from other two children households.

Currie and Yelowitz's IV strategy yields parameter estimates at odds with common perceptions about public housing. Public housing residents live in less crowded apartments in less densely populated buildings, and their kids are 11 percentage points *less* likely to be left behind than if the households had not lived in public housing. Currie and Yelowitz conclude that while public housing might serve its residents better if it placed them in less crowded units in better neighborhoods, in its current form, public housing does, on average, provide its residents with better housing and better opportunities for their kids than would they would get unassisted in the private housing market.

Test Scores and Class Size/Composition

Across the United States, billions of dollars are being spent to reduce class sizes. Are the gains in student performance worth the cost? Numerous other educational policies, for example desegregation plans that move children to schools outside their own neighborhoods, presume that peers can have strong effects on a child's educational outcomes. Is this presumption correct? Joshua Angrist and Victor Lavy (Angrist and Lavy 1999) and Caroline Hoxby (Hoxby 2000) ask whether smaller class sizes improve children's test scores. Hoxby (2002) asks whether peers affect children's test scores (Hoxby 2002). Angrist and Lavy study Israeli grammar schools; Hoxby examines Connecticut grammar schools.

OLS estimates of class size and peer effects are likely to be biased because local governments determine school sizes and parents determine much about who their children go to school with. The determinants of class size and peer group are quite apt to be correlated with unmeasured determinants of student performance. Parents willing to vote with their ballots and their feet to get small classes and more-desired peer groups are also apt to take other steps to improve their children's academic performance.

Hoxby exploits two instrumental variable strategies to consistently estimate the effects of class size and peers on children's test scores[9]. (One peer group effect she examines is the share of one's classmates who are female[10].) Her first IV strategy is to find empirically the unpredictable surprises in class size and class composition for each school in each year in her panel of Connecticut schools.

In small schools, year-to-year random fluctuations in local births can appreciably alter both class sizes and class compositions. Hoxby exploits her panel of schools to estimate forecasting models for the annual enrollment (or number of girls) in a cohort for each school. The instrument for class size or class composition is the prediction error for the given school in a given year. Hoxby reasons that because these increments to class size or peer group share are surprises, they are likely to be little correlated with the measures taken by parents to influence class sizes or choose one school over another[11].

When analyzing class size, Hoxby applies a second IV strategy as well. It is the strategy introduced by Angrist and Lavy. This strategy exploits variation in class sizes that is rooted in rules for maximum class sizes. When class sizes drop sharply only because a threshold class size has been reached and new smaller classes are, therefore, created, the change in class size is likely to be uncorrelated with other choices or circumstances of adults that would influence test scores. Examining classes of maximum size with smaller classes that result from the threshold being crossed provides a comparison free from the biases of OLS; this, in essence, is Hoxby's second IV strategy and the strategy employed by Angrist and Lavy. Like class size, some other rule-based variables, such as college financial aid (Van der Klaauw 1996), display similar "regression discontinuities" that can be exploited to construct instrumental variables.

Angrist and Lavy estimate that decreasing class size induces significant and substantial increases in Israeli children's test scores. Hoxby's analyses of Connecticut grammar schools estimate effects of class size that are small in magnitude, mixed of sign, and estimated with enough precision to reject even pretty small positive effects of decreased class size. As for peer effects, Hoxby finds that both boys and girls perform appreciably and statistically significantly better on reading and math tests when the share of girls in the class composition is higher.

Compulsory School and Earnings

Compulsory schooling requirements are commonplace in developed countries. Joshua Angrist and Alan Krueger (Angrist and Kreuger 1991) ask whether such

[9] Hoxby also uses fixed effects for each school in her panel of schools to control for omitted explanators that are fixed for each school.

[10] Hoxby (2002) also examines peer effects as reflected in the racial and ethnic compositions of classes. Those results do not lend themselves to simple interpretation, so I omit them.

[11] A few parents might react to an enrollment-day surprise in the size or composition of classes by not enrolling the child and sending the child to a private school, or even moving, instead. To drain even that small correlation from her instrument, Hoxby also looks at each cohort for each school and asks how "what was the surprise in the number of kindergarten-eligible children in the cohort in the year the cohort was eligible for kindergarten?"

requirements actually affect either educational attainment or earnings of American men. At issue is whether the requirements are binding, enforced, and efficacious. The key to the analysis is the observation that Americans born in different quarters of the year are affected differently by the attendance requirements.

In most American schools, students cannot enter the first grade unless they have reached age six by the end of the calendar year in which they enter school. These rules have the effect of making students born earlier in the year start school at a later age. Consequently, individuals born earlier in the year reach the age when they can leave school (16 or 17, depending on the state) with less education than individuals born later in the year, who tend to start school at a younger age. Angrist and Krueger convincingly document that in the U.S., males born earlier in the year do, indeed, tend to get less schooling than individuals born later in the year.

Angrist and Krueger note that the observed differences in education by quarter of birth provide the foundation for an instrumental variable for education in an earnings equation. The quarter of birth instrument is doubly interesting. First, it allows consistent determination of whether the extra education of men born later in the year garners those men higher wages. Second, because it provides an instrument that shields against omitted variables bias, it offers a check on whether the host of OLS-based earnings equations in the literature suffer from omitted variable bias.

Angrist and Krueger find that compulsory education regulations do lead men born later in the year to earn more than they otherwise would. Moreover, the estimated return to education obtained using the quarter-of-birth instruments proved similar to the OLS estimates (though somewhat larger), suggesting OLS does not suffer from much omitted variable bias.

The Intertemporal Elasticity of Substitution

For a broad class of preferences, called Epstein-Zin preferences, an investor consumes a constant fraction of wealth only if his or her elasticity of intertemporal substitution is one.[12] For a commonly assumed subset of those preferences, such a unitary elasticity also implies that the investor is myopic.[13] The intertemporal elasticity of substitution plays a further economically important role in many Neo-Keynesian macro models, where the elasticity is a parameter of the intertemporal IS curve that ties together the current interest rate, the expected future interest rate, and the equilibrium level of current output.[14]

Motohiro Yogo estimates the intertemporal elasticity of substitution, ψ, in each of eleven countries, and tests in each country the null hypothesis that the elasticity is equal to one (Yogo (2004)). For an Epstein-Zin utility maximizing consumer, ψ is the

[12] See Epstein and Zin (1989).
[13] Power utility functions constitute the narrower class of preferences. See Campbell and Veceira (2002) for details.
[14] See Woodford (2003)

slope coefficient in a regression of the current growth in consumption on an intercept and a current real rate of return. In the aggregate time series model that Yogo uses, real rates of return and changes in consumption are jointly determined. Consequently, OLS would suffer simultaneity biases if used to estimate ψ. Yogo uses lagged values of (i) the nominal interest rate, (ii) inflation, (iii) the log of the dividend-price ratio, and (iv) the growth in consumption as instruments. He estimates that the elasticity is small in magnitude in all eleven of the countries he studies and rejects everywhere the null hypothesis of a unitary elasticity.

Democracy and Education

Daron Acemoglu, Simon Johnson, James Robinson, and Pierre Yared (AJRY) repeat an oft asked, and oft answered question in political economy: Do higher levels of education in a country make democracy more likely (AJRY (2005a))? Their answer runs contrary to what political theorists expect and what other empirical analysts have found.

Many researchers have noted a strong positive correlation between education and indices of democracy. Consonant with these findings, Edward Glaeser, Rafael La Porta, Florencio Lopez-de-Silanes, and Andre Schleifer report that a fixed-effects regression of an index of democracy on the lagged value of the index of democracy, lagged education, and lagged income for a panel of countries yields a statistically significant positive coefficient on lagged education (Glaeser, at al. (2004)). AJRY extend Glaeser et al.'s approach two steps further – they add period-specific dummy variables to the model and perform IV estimation. IV estimation is needed because the presence of a lagged dependent variable in panel data ("a dynamic panel model") biases fixed effects estimation.

To obtain consistent estimates, AJRY follow the counsel of Arellano and Bond (1991). First, they estimate the model in first differences. Second, they use values of the level of democracy lagged two and more times as instruments for the change in the lagged democracy variable. Arellano and Bond show that this strategy is both consistent and efficient if there is no serial correlation in the disturbances of the model expressed in levels. AJRY are particularly interested in efficiency because they do not want findings of insignificant coefficient estimates to be due to great imprecision.

AJRY find no significant positive effect of education on democracy once they add period-specific effects to their regressions, with or without their various instrumental variables. The estimated standard errors in AJRY are, however, rather large. Although across numerous specifications the magnitudes of their various IV estimates are frequently negative, their 95% confidence intervals include positive values almost as large as estimated with OLS.[15] Thus, rather than empirically resolving the relationship

[15] In a related paper AJRY explore the relationship between income and democracy (AJRY (2005b)).

between education and democracy, AJRY's IV results yield a more a cautionary tale about the limits of our empirical knowledge of this relationship.[16]

Moving to Opportunity

Do poor households benefit from living in low-poverty neighborhoods? Do their kids? Jeffrey Kling, Jens Ludwig, and Lawrence Katz (KLK) address the latter question by examining data from the Moving to Opportunity experiment of the Department of Housing and Urban Development. In the experiment, a randomly selected sample of public housing tenants who applied for inclusion in the experiment were offered housing vouchers applicable to housing units in low-poverty neighborhoods. Not all selected households actually used the vouchers, but a substantial fraction did. Another random sample of public housing tenants who applied for inclusion in the experiment was not offered the voucher, but was tracked to provide a control group. KLK estimate the effect of the vouchers on arrest rates among youths.

KLK added covariates to the analysis to reduce the variance of the disturbances and thereby estimate more precisely the average effect on arrests of vouchers to housing in low-poverty neighborhoods for youths from households that used an offered voucher. They specified

$$A_i = \beta_0 + \beta_1 P_i + X\beta_2 + \varepsilon_i, \qquad (3)$$

where A_i is the number of arrests for the i^{th} youth, P_i is a dummy variable indicating whether the youth's household actually used a voucher, and X_i is a vector of covariates.

Whether a household uses its voucher (P_i) is likely to be correlated with ε_i because households decide whether to use a voucher based in part on what they expect the outcomes will be, and their expectations are likely to depend on relevant factors not measured in the study. An OLS estimate of β_1 is likely to be biased by this unobserved heterogeneity. A valid instrument for whether a youth's household used the voucher is a dummy variable indicating whether the youth's household was an experimental household. This is the instrument KLK use. They find that arrests for both violent crimes and property crimes drop markedly for female youths, and that arrests for violent crimes drop markedly for males, though perhaps only in the short run. Arrests for property crimes rise among the treated male youths.

IV estimation makes a critical difference in each of the nine papers just described. Levitt obtains estimates of incarceration's effects on crime that are 2-3 times those that OLS indicates. Currie and Yelowitz find large, statistically significant positive effects of public housing with IV estimation, which is opposite the suggestion of OLS. Hoxby uses IV to show that effects of class size on Connecticut students' performance thought statistically significant using OLS are statistically insignificant and quite small. Angrist

[16] Neither AJRY nor Glaeser, et al. treat education as endogenous. If one expects that more democratic societies spur education, then accounting for this endogeneity is likely to make AJRY's estimated coefficients even more negative, which would not much alter their conclusions.

and Lavy use IV estimation to convincingly show that the raw positive correlation between class size and achievement in Israel is due to smaller classes in Israel tending to have many disadvantaged students and that the actual effect of larger class sizes is to reduce Israeli students' achievement. Angrist and Krueger overcome the perils of omitted variable bias by relying on their quarter–of-birth instruments. AJRY use IV estimation to show that cross country panel data are not very informative about how education influences democracy. Yogo uses recently developed IV techniques to dramatically narrow the range of plausible estimates of the elasticity of intertemporal substitution. KKL exploit the MTO experiment's randomization to overcome unobserved heterogeneity in youths' responses to their neighborhood environment. All of the results are shielded against biases that would taint OLS estimates – or they are if the instruments the authors use are valid. How, then, do these authors argue for the validity of their instruments?

2. Avoiding Bad Instruments

Valid instruments must be uncorrelated with the disturbances in the regression of interest and the credibility of IV estimates rests on the arguments offered for the instruments' validity. This section describes nine strategies for evidencing the validity of instruments, and illustrates those strategies with the analyses described in the previous section. Seven of these nine strategies apply equally well to exactly identified and over-identified equations.

It might surprise some that there are so many strategies in support of a single identifying instrument. A common view is that when there is only one instrument for a troublesome variable, we can only rely on intuition for deciding whether an instrument is valid; "because [validity depends on] a covariance between [the instrument] and the error *u*, it can never be checked or even tested; we must maintain this assumption by appealing to economic behavior or gut feeling."[17] This view represents an advance over what was common practice twenty years ago and more, when IV regressions often included everything but the kitchen sink as instruments – if a variable was in one's data set and not in the equation in question, it would be tossed in among the instruments; see Murray (1983) for a vintage example. But we can, in fact do more than just rely on our intuition, even when an equation is exactly identified.

The strategies discussed in this section are:

i. Test Over-identifying Restrictions
ii. Check for Serially Correlated Disturbances
iii. Use Alternative Instruments
iv. Preclude Links Between the Instruments and the Disturbances
v. Use Information from Other Populations
vi. Be Diligent About Omitted Explanators
vii. Randomize

[17] Wooldridge (2000), p. 463. Wooldridge abandons this language in subsequent editions, but the quote captures well what I think many would say.

viii. Use Economic Theory
ix. Use Intuition and Reduced Forms

These strategies can each exclude one or another rationale for an instrument's being invalid.

Test Over-identifying Restrictions

One long-standing strategy for assessing instruments' validity applies when the number of instruments exceeds the number required for exact identification, that is, when $l > q$. For example, one could use exactly q instruments and 2SLS to estimate equation (1) with the remaining instruments added to equation (1) as possible explanators. Rejecting the null hypothesis that these remaining instruments all have zero coefficients would reject the validity of one or more of these remaining instruments, conditional on the validity of the q instruments used in 2SLS. When an equation is over-identified, the over-identifying exclusion restrictions should be tested. This has been standard practice for a long time.

Several econometric theorists (for example, Anderson and Rubin (1949), Hahn and Hausman (2002), Hansen (1982), Hausman (1983), and Sargan (1958)) have proposed formal statistical tests for the validity of instruments. The formal over-identification tests all need, implicitly or explicitly, consistent estimates of equation (1) – hence they all need at least enough valid instruments to exactly identify the equation before they can provide tests of the validity of other instruments. It is this inherent limitation of such formal tests that leads many to think that with only a single instrument, only intuition or theory can serve us.

Stephen Levitt's crime rate equations are potentially over-identified. His instruments include several changes in the status of prison over-crowding lawsuits, such as filing and preliminary decision, and they distinguish between filing changes in the year of an observation and filing changes in years preceding an observation. In all, this yields ten instrumental variables for the one troublesome variable. Levitt uses Sargan's test of over-identifying restrictions to assess his instruments.[18,19] Conditional on at least one of his instruments being valid, Levitt fails to reject the null hypothesis that his instruments are valid.

Like Levitt's crime rate equation, AJRY's democracy equation is potentially over-identified. When estimating the effect of changes in education on changes in democracy from one five-year period to the next, AJRY use several multiply-lagged values of democracy as instruments.

---

[18] Sargan's test statistic is nR^2 using the R^2 from a regression of residuals from equation (1) fit using the 2SLS estimates of the parameters on the elements of Z. The statistic has a Chi-square distribution with degrees of freedom equal to $(l\text{-}q)$, the degree of over-identification.

[19] The Stata command ivreg2 (Baum et al., 2003) yields Sargan's test statistic. This command is an add-on to Stata. To locate the ivreg2 code from within Stata, type "findit ivreg2" on Stata's command line. Then click on the website name given for ivreg2 to update Stata. In Eviews, the GMM procedure reports Hansen's J-test, a more general version of Sargan's test.

AJRY test their over-identifying restrictions; they use a test due to Hansen. When there are multiple instruments for a variable, 2SLS provides the optimal combination of those instruments if the disturbances in equation (1) are serially uncorrelated and homoskedastic for all values that the explanators can take on. When the disturbances are heteroskedastic or serially correlated, 2SLS is no longer efficient; the generalized method of moments (GMM) estimator uses multiple instruments more efficiently in this case. AJRY use GMM to estimate the relationship between changes in democracy and changes in education. To test their over-identifying restrictions, AJRY apply Hansen's *J*-test, which is, in essence, a generalization of Sargan's test to the GMM context.[20] The Hansen test does not reject the over-identifying restrictions.

A failure to reject the over-identifying restrictions in a model lends credibility to IV estimates. However, when sample size is small or instruments are weak, the nominal significance level of Sargan and Hansen's tests are well below the actual significance levels in practice – the tests reject the null hypothesis of valid instruments too often. This "size distortion" in these tests makes rejections of the null hypothesis less informative about the validity of instruments.

Tests of over-identifying restrictions are most compelling when there are some instruments (enough to identify the equation) whose validity seems sure. It is then that we can be confident the test is itself valid. Tests of over-identifying restrictions are particularly suspect when all of the instruments share a common vulnerability to being invalid. For example, in Levitt's crime study, all of the instruments are linked to over-crowding lawsuits. If one lawsuit-related instrument is invalid, we are apt to worry that they all are – and therefore that Sargan's test is invalid. In contrast, perhaps we are confident in AJRY's study that the levels of democracy at long lags are uncorrelated with the disturbances, and our sole worry is whether the levels at shorter lags are actually uncorrelated with the disturbances. If so, then we can also be confident that AJRY's over-identification test is valid, because the longer lags of democracy suffice to identify the equation.

Some economists are very wary of over-identification tests because they rest on there being enough valid instruments to exactly identify the relationship. Their worry is that too often a failure to reject the null of valid over-identifying restrictions tempts us to think we have verified the validity of *all* of the instruments. We should discipline ourselves not to succumb to that temptation.

Check for Serially Correlated Disturbances

In time series data, lagged dependent variable explanators are likely to be correlated with the disturbances if the disturbances are autoregressive, but to be free of such correlation asymptotically if the disturbances are not serially correlated. Testing for serial correlation in models with lagged dependent variables is a useful check for whether

[20] Eviews and Stata's ivreg2 include options for using GMM and for conducting Hansen's *J*-test.

or not IV estimation is needed. When OLS is consistent, it is more efficient than IV estimation; we don't want to use instrumental variables needlessly.

In fixed or random effects panel data with a large number of cross sections and relatively few time periods (as is often the case with panel data), a lagged dependent variable among the explanators makes least squares estimators inconsistent even when the disturbances are not serially correlated. An IV solution is needed in these cases.

The dynamic fixed-effects model estimation procedure recommended by Arellano and Bond (1991), and used by AJRY in studying the link between education and democracy, is consistent if the disturbances in the model are not first order autocorrelated. It is also efficient. Arellano and Bond offer a test for serial correlation that is applicable in this case. Their test is also applicable more generally. AJRY use this test to check that their disturbances are not serially correlated. They fail to reject the null of no serial correlation in the disturbances in their dynamic panel model of democracy.[21] Showing that their disturbances are not serially correlated enhances the credibility of AJRY's IV estimates.

Use Alternative Instruments

Over-identification tests formally ask, in essence, whether all of the instruments tell the same story about the parameters of interest. 2SLS facilitates this formal process by providing a strategy for using all instruments in a single estimation procedure. Sometimes, however, it is not feasible to include all instruments in one estimation. In such cases, there is still information to be had by comparing the results from applying several instruments separately from one another. If the parameter estimates using different instruments differ appreciably and seemingly significantly from one another, the validity of the instruments becomes suspect. If all of the estimates are consonant with a single interpretation of the data, their credibility is enhanced.

Caroline Hoxby uses two distinct identification strategies in her study of the effects of class size on test scores. In one, she uses surprises in enrollments; in the other, she relies on maximum class-size rules. In the former, her units of observation are schools in a given year. For the latter, she compares differences between cohorts in a given school in years just before and just after a class-size rule triggered a change in class size, and looks only at data from such events. These two IV strategies do not lend themselves to a single estimation procedure with which formal over-identification tests could be applied. Nonetheless, it boosts confidence in the two sets of IV results that both indicate very small effects of class-size on performance.

Preclude Links Between the Instruments and the Disturbances

[21] The Arellano-Bond test does not require full-blown GMM estimation. It is also applicable with OLS and 2SLS estimation (both of which are special cases of GMM). The command "findit abar" in Stata will reveal the location of the code for the Arellano-Bond test. Click on the location address to upgrade Stata with the abar procedure.

Thoughtful critics of your favorite IV analysis will conjure reasonable stories to explain why your instruments might be invalid. Anticipate these criticisms. Test the stories. Doing so makes your results more credible.

Steven Levitt anticipated that critics would argue that prison overcrowding lawsuits might result from past swells in incarceration that arose from unusually high past crime rates. If this were so, and crime rates were serially correlated, Levitt's instruments would be invalid. Levitt tackled the possibility head-on. He asked whether over-crowding lawsuits could be predicted from past crime rates, and found they could not. By deflecting one challenge to the validity of his instrument, Levitt enhanced the credibility of his results. Notice that Levitt's strategy did not require that his equation be over-identified. It would have worked for even a single overcrowding lawsuit variable.

Caroline Hoxby anticipated that some might worry that parents in her sample reacted to finding a surprisingly large class by not enrolling their child in the local school, but turning instead to a private school or moving to another school's area. She blunted this criticism by using an alternative instrument to surprises in enrollment. In addition to using surprises in enrollment, Hoxby used surprises in the size of a cohort when the cohort reached kindergarten age. She thereby included in the instrument those children who were subsequently sent to other schools. Because the kindergarten-eligible-cohort surprises and the enrollment surprises are highly correlated, Hoxby was probably wise not to use both at once. The mean square error of IV estimation can rise with the number of instruments; highly correlated instruments should be avoided. Because both instruments led to the same quantitative conclusions, Hoxby's results were buttressed.

Josh Angrist and Alan Kreuger anticipated worries that quarter of birth might be correlated with unobserved variables relevant to earnings, and that it was the effect of these, not of compulsory schooling that the authors were detecting. To counter this concern, Angrist and Krueger asked whether high school graduates (who are exempted from compulsory schooling laws) born earlier in the year ultimately had less education than other high school graduates. The answer was basically no - post-secondary educational attainment is not less for students born earlier in the year. By implication, quarter of birth is not correlated with unobserved variables that influence educational attainment, unless those variables only influence primary or secondary school attendance. Angrist and Krueger also include the quarter-of-birth variables in two earnings equations, one for the whole sample and including education as an explanator, the other for only college graduates and with only quarter-of-birth dummies as explanators. If the quarter of birth is correlated with anything relevant to earnings besides compulsory attendance, the quarter-of-birth dummies should be significant in such regressions. They are not, which lends strong support to the validity of the quarter-of-birth instruments.

Use Information from Other Populations

An instrument is not valid if it is an omitted explanator in the model. Establishing that one's instrument does not itself belong among the explanators in the equation would seem especially difficult in the case of a single instrument. A variable can't be both an

explanator and an instrument for another variable; just as with perfect multicollinearity and OLS, including a variable in both of these roles makes IV estimators not exist.

Janet Currie and Aaron Yelowitz tackle doubts about whether their instrument should be an explanator in their study of public housing and kids' school outcomes. Their instrument is the sex mix of a two-child family: same-sex or different-sex. The potential challenge to the validity of their instrument is that perhaps the academic outcomes of kids in two-parent households depend on the sex mix of the family, and not just on the sex of the child. Perhaps their instrument should be an explanator.

Currie and Yelowitz answer this challenge by turning to the psychology literature on children's achievements. The vast majority of households in those studies are neither in, nor even eligible for, public housing. Consequently, the fact that those studies omitted public housing participation in studying the determinants of kids' success almost surely does not matter for their parameter estimates. What Currie and Yelowitz conclude is that the literature is in agreement that sex mix does not matter for boys, and might, but probably doesn't, matter for girls. Thus, their instrument is not an excluded explanator when studying academic outcomes for boys, and is probably not an excluded explanator when studying girls. Currie and Yelowitz buttress their IV findings, especially those for boys, by turning to the data for another, closely related, population.

Josh Angrist and Alan Krueger also look to evidence from other populations. They report studies that find no correlation between quarter-of-birth and the traits of one's parents. This further supports Angrist and Krueger's claim that their instrument is not correlated with relevant omitted variables.

Data from other sources is as relevant for exactly identified IV estimations as for over-identified IV estimations. The primary caution when looking to other populations for information about one's instruments is that populations often do differ from one another. The force of the external information is lessened the more different the population is from the population sampled in your data.

<u>Be Diligent About Omitted Explanators</u>

An instrument is not valid if it is correlated with an omitted relevant variable. Even if an omitted variable is uncorrelated with the troublesome explanator, it can bias IV estimation if it is correlated with the instrumental variable. This concern requires that we be doubly vigilant about omitted variables when doing IV estimation. We are accustomed to thinking abut what variables are omitted when we consider OLS, because omitting a relevant variable biases OLS if the included explanators are correlated with the omitted variable. We ask, "Is X correlated with things I am omitting." Sometimes we can convince ourselves that we have included the potential explanators with which X is most likely to be correlated. Turning to IV estimation, we must ask the question again. Are there any omitted explanators that might be correlated with my instrument, even ones that we don't believe are correlated with X?

AJR are careful about what variables they include in their IV estimation of the relationship between economic growth and the risk of property expropriation. In their first OLS and IV specifications, they do not include current health conditions among the explanators for income. Arguably, current health conditions are not be correlated with the risk that private property will be expropriated, the explanatory variable of interest. Omitting these health variables from an OLS regression might not be troubling. Omitting these current health variables is very worrisome, however, when the instrumental variable is death rates at the time of colonization. Thoughtful critics are likely to worry that past health conditions and present health conditions are correlated. Such a correlation would bias IV estimation if the current health variables were omitted from the regression. AJR note this concern and therefore include measures of current health conditions in their final model. When AJR add a malaria index for 1994 to their model, 2SLS continues to yield a statistically significant positive coefficient on the risk of expropriation. This strengthens confidence in their results. However, when both continent dummies and the malaria index are both added to this same specification, 2SLS finds an even larger, but statistically insignificant effect of the risk or expropriation.[22] Thus, careful attention to what variables to include in the model makes us more cautious about the lessons to draw from these data. (However, below we find further reason to think that these data do, on balance, indicate a relationship between the average risk of expropriation and growth.)

Levitt is careful about potential omitted variables in another study of crime rates that investigates the relationship between changes in crime rates and changes in the number of police officers per capita (Levitt (1997)[23]). Levitt's instrument is mayoral election cycles. He is careful to include among his explanators changes in local welfare and education expenditures. While the correlation between numbers of police and welfare or education expenditures might be small, their correlation with mayoral election cycles is likely to be large. Omitting such variables could seriously bias IV estimates of the effect of police officers on crime rates.

Including those variables most likely to be correlated with an instrument builds the credibility of IV results in both exactly identified and over-identified equations.

Randomize

Randomized experiments are the gold standard for scientific experimentation. The best the rest can hope for is the status of "quasi-experiment" – an effort that takes steps to overcome some of the hazards of unrandomized experiments, but that, in the end, risks unknown biases.[24] Most instrumental variables, in the end, define quasi-experiments. However, sometimes, a situation offers a randomly assigned instrumental variable. In these cases, if the estimated equation omits no endogenous variables, and if the randomly assigned variable is not itself a variable in the equation being estimated, the lack of correlation between the instrument and the disturbances is guaranteed.

---

[22] I am grateful to Daron Acemoglu and Simon Johnson for sharing their data with me.
[23] McCrary (2002) and Levitt (2002) should be read in conjunction with Levitt (1997).
[24] Even randomized experiments can be undermined by refusals to participate or by attrition.

KLK's IV estimates of the effects of using a voucher to move to a low-poverty neighborhood on poor youths' arrest rates are credible because the instrument, whether one's family is an experimental or a control household, is randomly assigned. It is plausible that neither control households, nor experimental households who do not use the voucher, are affected by being in the experiment, so experimental status is not a variable in the equation of interest.[25]

Use Economic Theory

Unlike statisticians, econometricians readily rely on economic theory in formulating their empirical analyses. Economists have considerable confidence that their theories tell us about how the world works; it makes sense that we incorporate this knowledge into our empirical work. In particular, economic theory can sometimes tell us that an instrument is valid. For example rational expectations and the efficient market hypotheses declare that current changes in some variables will be uncorrelated with all past outcomes. In such cases, lagged variables can serve as instruments for those changes.

Yogo justifies using lagged economic variables as instruments in just this way. Robert Hall wrote "Actual movements of consumption differ from planned movements by a completely unpredictable random variable that indexes all the information available next year that was not incorporated in the planning process the year before" (Hall 1988). Because rational consumers would incorporate into their planning the economic variables known in the year when plans were made, lagged economic variables will be uncorrelated with changes in consumption in the current year. To overcome any problems raised by actual consumption measures being aggregated across a year, Yogo lags his instrumental variables two years, instead of one.

When an economic theory relied on to justify instruments has been tested and found correct in other settings, using economic theory becomes a sophisticated way to appeal to information from other populations. When an economic theory relied on to justify instruments has not been empirically tested, using economic theory becomes a formal version of the last category for buttressing an instrument's validity – using intuition.

Use Intuition and Reduced Forms

Steven Levitt argues (p. 323) that his litigation status instruments are valid because "it is plausible that prison overcrowding litigation will be related to crime rates only through crime's impact on prison populations, making the exclusion of litigation status itself from the crime equation valid." An intuitive argument for why an instrument is valid is better than no argument. Indeed, an unintuitive instrument requires considerable justification if people are to accept your IV results. When there is an intuitive rationale for an instrument's validity, people who share that intuition will have reason to give credence to the IV results. When intuition is buttressed by other arguments, like those noted above, the credibility of IV results is further enhanced.

[25] An early, classic use of a randomized instrument appears in Angrist (1990).

Of course, intuition need not stand naked and alone. Intuition can be checked. Reduced-form regressions with the instrumental variables and the non-troublesome explanators as the explanatory variables, and either the dependent variable of interest or the troublesome explanatory as the dependent variable are unbiasedly estimated with OLS if the instruments are valid. These reduced form regressions provide valuable information with which to check one's intuitions. Angrist and Krueger (2001) point out (p.80) that if the candidate instruments are all statistically insignificant in the reduced form equation for the dependent variable of interest, the presumption should be that either the model is under-identified and IV estimation is uninformative or the troublesome variable does not matter for the dependent variable of fundamental interest. Moreover, finding that an instrumental variable appears in the reduced form equations for either the troublesome variable or the dependent variable of interest with a sign that is at odds with the instrument's intuition requires rethinking the instrument's validity. In the extreme, if no instruments appear in the reduced form equation for the troublesome variable – if all the instruments are irrelevant to the troublesome variable - there is no valid instrument. Always exploit the valuable information to be had from examining the reduced forms. [26]

Angrist and Lavy, in their regression discontinuity IV analysis of the effect of class size on students' test scores, give considerable attention to the reduced form relationships for actual class size and test scores, with particular attention to the rule-based class-size prediction that the authors use as an instrument. The authors note that "(t)he reduced-form relationship between predicted class size (f_{sc}) and actual class size reported in Table II for a variety of specifications, shows that higher predicted class sizes are associated with larger class sizes and lower test scores." (p.552) Angrist and Lavy then discuss their reduced form results with an eye to substantiating their claim that their instrument is a good one. Angrist and Lavy also offer a related graphical strategy. First they show graphically that class sizes do move with enrollments in much the discontinuous fashion that Israeli rules call for. They then show graphically that average test scores also move with discontinuities that roughly follow the discontinuous rule-predicted class sizes. The Israeli data support Angrist and Lavy's identification story. Angrist and Krueger use a similar graphical strategy to support their intuition that educational attainment is linked to quarter of birth. A graph makes clear that in the paper's 30-year sample years of education tends to rise with quarter of birth.

Levitt (1996) considers regressions with the prison-overcrowding litigation instruments as explanators for his dependent variable (changes in crime rates). Levitt finds that in this regression, the instrumental variables all have coefficients that are significantly different from zero with signs that support his identification story: increases in crime rates follow litigation, especially successful litigation. Furthermore, Levitt finds that the litigation variables for the period just *before* litigation are associated with

[26] Pre-testing variables in regression analysis has long been known to lead to inconsistency (Leamer (1978), Miller (1990)) More recently, Hansen, Hausman, and Newey (2005) explore pre-testing in the specific case of IV estimation; fishing among individual instruments to find significant ones is a poor idea. The set of instruments should be assessed together. Arellano, Hansen, and Sentana (AHS) offer a formal test for the relevance of a set of instruments (AHS (1999)).

increases in prison populations, while the litigation variables for the period during the litigation and after a judgment unfavorable to the state are associated with declines in prison populations, as his identification story would suggest.

AJR examine the performance of their instrument, the early European settler mortality rate, in the reduced form equations. They reject the null hypothesis that their one instrument has a zero coefficient in the reduced form equation for the risk of expropriation, and they find that the settler mortality rate coefficient is significant and negative in the reduced form equation for income, as their identification story predicts.

Currie and Yelowitz first analyze the effects of public housing using data from the Survey of Income and Program Participation (SIPP). They find that the t-statistic on their one instrument, the sex-mix of siblings, is insignificant. Their instrument seems to be irrelevant. However, there are several reasons why we might reject the null hypothesis of no effect of an instrument on a troublesome variable. No effect is one; a small sample size is another. Currie and Yelowitz concluded that the 86 participants in public housing in their SIPP sample were simply too few to see the efficacy of their instrument. Undaunted, they shifted their attention to the Current Population Survey, which provided many more observations on households in public housing. In these data, sex-mix was significant.

Pre-testing variables in regression analysis has long been known to lead to inconsistency (Leamer, 1978; Miller, 1990). Recently, Hansen, Hausman, and Newey (2005) explore pre-testing in the specific case of instrumental variable estimation; they conclude that fishing in a set of potential instruments to find significant ones is also a poor idea. The set of instruments should be assessed together -- Arellano, Hansen, and Sentana (1999) offer a suitable formal test. With data mining frowned upon, it is all the more important to diligently apply intuition when selecting potential instruments.

IV Estimation Is Not a Panacea

IV estimation can cure so many ills that we are tempted to think of it as a panacea. Find a promising instrument and our work is done. But our work is not done so simply. All instruments arrive on the scene with a dark cloud of invalidity hanging overhead. This cloud never goes entirely away, but researchers should chase away as much of the cloud as they can. That OLS can suffer biases from multiple sources complicates dispelling the clouds of invalidity. For example, repeated measurement of an explanator can provide an instrument that overcomes the bias OLS suffers in the face of a mis-measured explanator, but such an instrument might nonetheless be invalid if the mis-measured explanator is also endogeneous. And instruments valid against measurement error and endogeneity might still be invalidated by omitted relevant explanators.[27]

The truly scientific value of the strategies outlined in this section is that several of them can formally expose an invalid instrument. Had AJRY's over-identification test

[27] See Angrist and Krueger (2001) for a lucid discussion of the relationships among omitted variables, natural experiments, and using IV estimation to estimate causal relationships.

rejected their over-identifying restrictions, or if their disturbances had displayed first-order serial correlation, their reliance on Arellano and Bond's procedures would have become futile. If Hoxby had obtained sharply different IV estimates from her two instruments, or Levitt had found that crime rates predicted over-crowding litigation, or Currie and Yelowitz had found that the sex-mix of siblings mattered for kids' outcomes among non-poor households, their instruments would have been dismissed.

Others of the strategies, most notably diligence about omitted variables and intuition, only serve to thin the clouds of invalidity – they offer no formal criterion for rejecting an instrument for its invalidity. In the end, if all tests are passed, prognostication about an instrument's validity based upon the remaining clouds is a subjective matter. How much credence you grant any one of the empirical studies discussed here may differ from how much credence I grant it. Indeed, among colleagues who have read the papers discussed here, I find a range of views about how much credence to give these paper's instruments. But what we can all agree on is that we would each be more skeptical of each of these studies had the authors not subjected their data to one or more of the assessments discussed in this section.

3. Coping with Weak Instruments

Relevant instruments are correlated with the troublesome variable. Irrelevant instruments are of no use in estimation. When valid instruments are strongly correlated with the troublesome variable, 2SLS is an effective tool. When instruments are weak, that is, weakly correlated with the troublesome variable, however, 2SLS loses its attractions.

When instruments are weak, researchers should forego 2SLS and turn to alternative procedures for testing and estimation. Some econometric theorists even argue that researchers should always forego 2SLS in favor of alternative IV methods (Andrews, Moreira, and Stock (2005) and Andrews and Stock (2005)). If one is planning on using 2SLS, one should first rule out weak instruments. This section reviews old and new practices that ensure an analysis is unlikely to succumb to the pitfalls of weak instruments.

The Virtues of Strong Instruments

The statistical virtues of strong instruments have long been understood. If equation (1) is over-identified ($l > q$) and the number of instruments is not large relative to the sample size, strong instruments generally make the finite-sample biases of 2SLS small and inferences based on 2SLS's asymptotic normal distribution and its estimated standard error approximately valid (that is, the nominal sizes of tests are approximately the true size) in moderately large samples. Even in the exactly identified case ($l = q$), when the finite-sample mean of 2SLS doesn't exist, the estimator's median is equal to the true parameter value, and inferences based on 2SLS tend to be approximately valid in moderately large samples.

The virtue's of 2SLS are always partially offset, and are sometimes overwhelmed, by 2SLS's drawbacks. The non-existence of 2SLS's mean in the exactly identified case, and the non-existence of its variance when (l - q) = 1, mean that 2SLS can be wildly wrong more often than we might anticipate.[28] And despite its consistency, 2SLS is always biased in finite-samples.

2SLS's Finite-Sample Bias

A look at the finite sample bias of 2SLS points to why weak instrument can undermine the attractions of 2SLS. Simplifying equations (1) and (2) and choosing convenient units of measure for Y_1 and Y_2 highlights the problems and loses no substance.

Assume there are no untroublesome explanators in equation (1), so that $\beta_2 = 0$ and $\alpha_2 = 0$. Choose units of measure for Y_1 and Y_2 such that Var(ε_i) = 1 and Var(μ_i) = 1. A consequence of these variance assumptions is that the Cov(ε_i, μ_i) equals the correlation coefficient of ε_i and μ_i, which we call ρ. Because the instruments in Z are uncorrelated with ε_i, ρ also measures the degree to which Y_2 is troublesome, that is the degree to which it is correlated with the disturbances in (1). Finally, call R^2's population analog for the reduced form equation $\tilde{R}^2$.[29]

Jinyong Hahn and Jerry Hausman show that, in this specific simplified specification, the finite-sample bias of 2SLS when $l > q$ is, to a second order approximation,

$$E(\beta_1^{2SLS}) - \beta_1 \approx \frac{l\rho\left(1-\tilde{R}^2\right)}{n\tilde{R}^2} \qquad (3)$$

when the instruments are valid (Hahn and Hausman (2002)).[30] Consonant with the consistency of 2SLS, this bias goes to zero as n grows. This is unsurprising. More attention-getting is the role of $\tilde{R}^2$.[31] The smaller the population R^2 from Equation (2), the larger the finite-sample bias of 2SLS. A weak instrument can cause a large finite sample bias for 2SLS. With a very weak instrument, 2SLS might be seriously biased in even quite large samples.

---

[28] The nub of the problem is that the sample covariance between the instrument and the troublesome variable - and therefore the numerator of the 2SLS estimator - can get very close to zero more easily than can the variance of the troublesome variable that appears in the denominator of the OLS estimator.

[29] Nelson and Startz (1990a,b) first showed that much could be learned about the small sample properties of IV estimation by focusing on such a simple case.

[30] For earlier, more general renderings of 2SLS's finite sample properties when instruments are valid, see Rothenberg (1983, 1984) or Phillips (1983). When instruments are weak, second order approximations can prove to be poor approximations, but the substantive concern remains the same (see, for example, HHK (2004)).

[31] The role of l is also worth noting. Many instruments make for a larger finite-sample bias in 2SLS. Many instruments also bias downward the estimated standard error of 2SLS; when l is large relative to n, 2SLS suffers from a size distortion. This paper does not discuss methods for coping with many instruments; see Hansen, Hausman, and Newey (2005) for one useful strategy.

Hahn and Hausman also show that the ratio between the finite-sample biases of OLS and 2SLS in this specification is

$$\frac{Bias\left(\beta_1^{2SLS}\right)}{Bias\left(\beta_1^{OLS}\right)} \approx \frac{l}{n\tilde{R}^2}. \qquad (4)$$

As long as $n\tilde{R}^2$ is larger than the number of instruments, 2SLS has a smaller bias than OLS. Furthermore, because the right hand side of (4) is always positive, 2SLS tends to be biased in the same direction as OLS when there is a single troublesome variable.

Because the population analog of the *F*-statistic for equation (2) is

$$\tilde{F} = \frac{n\tilde{R}^2}{l\left(1-\tilde{R}^2\right)},$$

the approximate bias of 2SLS can also be expressed as:

$$Bias\left(\beta_1^{2SLS}\right) \approx \frac{1}{\tilde{F}} \frac{1}{1-\tilde{R}^2} Bias\left(\beta_1^{OLS}\right). \qquad (5)$$

Thus, when the $\tilde{R}^2$ in the reduced form equation is low, $\tilde{F}$ must be markedly greater than one if the bias in 2SLS is to be substantially lower than that of OLS. This observation suggests that when the R^2 of the estimated reduced form equation is low, we should look to the *F*-statistic of the reduced form equation to determine whether valid instruments are, collectively, strong enough that the bias in 2SLS is likely to be small relative to the bias of OLS. James Stock and Motohiro Yogo provide critical values for just such a test (Stock and Yogo (2005)). When there are untroublesome explanators in equation (1), the appropriate *F*-statistic is that for the null hypothesis that the parameters on all of the instruments (the variables in *Z*) are zero in equation (2).

Charles Nelson and Richard Startz show that the estimated variance of 2SLS is generally biased downward in finite samples (Nelson and Startz (1990b)) and that the bias can become quite large when the instruments are weak. Thus, weak valid instruments are likely to distort the size of tests based upon 2SLS - null hypotheses are too often rejected because the estimated variances are too small. Unfortunately, unbiasedly estimating the standard errors of the 2SLS estimator is not enough to obtain valid inferences from 2SLS when instruments are weak.[32] Stock, Yogo, and Jonathan Wright (SWY) show in a survey of weak instruments research that weak instruments make the asymptotic normal distribution of 2SLS a poor approximation to its finite-sample distribution (SWY (2002)). It is important to note that the weak instruments

[32] Improving the estimates of the standard errors can, however, overcome the size distortions that arise in 2SLS when the number of strong, valid instruments is large. See Hansen, Hausman, and Newey (2005).

problem is not just a small sample problem. Instruments can be weak in extremely large samples (Staiger and Stock (1997)).

How, then, are we to cope with weak instruments if they seriously bias 2SLS and seriously distort hypothesis tests based on 2SLS?

A Formal Test for Weak Instruments

For analysts who wish to continue using 2SLS when that is not bad practice, the first step is determining whether the available instruments are weak. The reduced-form *F*-statistic for valid instruments can tell us when 2SLS supports relatively unbiased estimation and relatively valid statistical inference. Stock and Yogo have computed the appropriate critical values for this *F*-statistic. 2SLS is a poor estimation choice when the instruments' strength is in doubt. Therefore, before using 2SLS, always conduct a Stock – Yogo test.[33] This test's use of an *F*-statistic highlights that weakness of an instrument is less about individual instruments, per se, than about the actual instrumental variable we construct from those individual components. The *F*-statistic used in a Stock-Yogo test pertains to only the set of instruments – the elements of Z - in the first stage of 2SLS. The test statistic is the *F*-statistic for the null hypothesis that the coefficients on the elements of *Z* are all zero. When elements of equation (1)'s *X* vector are included in the reduced form equation, the Stock-Yogo weak instrument test does not constrain their coefficients, only the coefficients of the elements of Z. One instrument weakly correlated with the troublesome variable is not likely to lead to a weak instrument when there are other instruments strongly correlated with the troublesome variable.

If the Stock-Yogo test rejects the null hypothesis that the instruments are weak, 2SLS estimates are probably not much biased and inference based on 2SLS is probably valid. I say "probably" because pre-testing for weak instruments changes the distribution of the 2SLS estimates one examines. Stock and Donald Andrews suggest foregoing 2SLS altogether because they prefer to avoid the potential pitfalls of pre-testing (Andrews and Stock (2005)). Nonetheless, long-time users of 2SLS may prefer to use 2SLS when the Stock-Yogo test indicates that their instruments are strong.

The theoretical *F*-value ($\tilde{F}$) required for valid inference with 2SLS is larger than that required for unbiased estimation. Consequently, there are two groups of critical values in a Stock-Yogo test. The first group applies for testing the null hypothesis that the true significance level of hypothesis tests based on 2SLS is below 10%, 15%, 20%, or

[33] Cruz and Moreira (2005) note that the power of the Stock-Yogo test varies from case to case and can be low. Rejecting the null hypothesis of weak instruments is generally reliable, but a failure to reject weak instruments can sometimes reflect low power of the test. Cruz and Moreira find indications of the variable power of the Stock-Yogo test in Angrist and Kreuger's compulsory education paper. In some of that paper's 2SLS regressions with instruments that are weak according to a Stock-Yogo test, 2SLS-based confidence intervals are virtually the same as corresponding confidence intervals built with methods robust to weak instruments, but for other such regressions, the 2SLS intervals are moderately narrow even though the robust intervals are infinitely wide. Since one can't determine which case one's own results fall into, 2SLS should be avoided and robust procedures used when a Stock-Yogo test does not reject the instruments being weak.

25% when the nominal level is 5%. The second group of critical values applies for testing the null hypothesis that the bias of 2SLS is greater than 10%, 15%, 20%, or 30% of the bias of OLS. The critical values depend on the number of instruments, *l*.

Table 1 replicates a subset of the critical value tables from Stock and Yogo (2005). These correspond to a null hypothesis of greater than 10% of the OLS bias and a true size of greater than 10%. Notice that the critical values when testing for undistorted size are much larger than those applicable when testing for relative unbiasedness, and they also rise more sharply with the number of instruments. When there are multiple troublesome variables, the instruments appear in several reduced form equations. The appropriate test statistic is not an *F*-statistic, but a Cragg-Donald statistic, which is the multiple equation analog of the *F*-statistic.[34] See Stock and Yogo (2005) for the additional critical values.

TABLE 1
SOME CRITICAL VALES FOR THE STOCK-YOGO TEST
OF THE NULL HYPOTHESIS THAT INSTRUMENTS ARE WEAK

	q=1		*q=2*	
	bias >10% of OLS bias	*true signific. Level > 10% when nominal level is 5%*	*bias >10% of OLS bias*	*true signific. Level>10% when nominal level is 5%*
l				
1	.	16.38	.	.
2	.	19.93	.	7.03
3	9.08	22.30	.	13.43
4	10.27	24.58	7.56	16.87
5	10.83	26.87	8.78	19.45
6	11.12	29.18	9.48	21.68

l is the number of instruments; *q* is the number of troublesome variables. When *q* = 1, the test statistic is an *F*-statistic. When *q*>1, the test statistic is the Cragg-Donald statistic.
Source: Stock and Yogo (2005)

Yogo compares the first stage *F*-statistics in his 2SLS estimation of the intertemporal elasticity of substitution to the Stock-Yogo critical values. He finds that in his quarterly data, the real aggregate stock return is so poorly predicted by the reduced form equation that he fails to reject the null hypothesis that 2SLS estimates will be seriously biased, nor does he reject the null hypothesis that 2SLS-based hypothesis tests are seriously size distorted. He concludes that 2SLS is inappropriate for estimating his model with stock return data. Real interest rates, however, prove predictable enough that 2SLS provides relatively unbiased parameter estimates, but not undistorted t-tests.

The other applied papers summarized in this article did not have the benefit of Stock and Yogo's critical values to check for weak instruments, but many heeded the

[34] The Cragg-Donald statistic is available as an option in Stata's ivreg2.

earlier counsel of Stock and Douglas Staiger to always check the first-stage *F*-statistic for the instruments being zero to ensure that it is "large" (Staiger and Stock (1997)). Prior to Stock and Staiger, few papers reported the results of first-stage regressions of 2SLS; six of the nine papers reviewed here report first-stage results, though only five of those report the needed statistic.

In her study of test scores, Hoxby reports a first-stage t-statistic in excess of 10 (*F*-statistic of 100) for her instrument based on the surprises in a cohort's size when eligible for kindergarten, and an even higher reduced form t-statistic for her enrollment surprise instrument. Her instruments are not weak. In asking whether public housing leads to good outcomes for kids, Currie and Yelowitz report a reduced form t-statistic of 4.14 on their sex-mix instrument in the CPS data. Their instrument is not weak.[35] In their study of education and democracy, AJRY do not report first-stage regression results.

In his study of crime rates and incarceration, Levitt reports reduced form p-values less than .001 for the *F*-test of zero coefficients on all of his instruments, but with 10 instruments, that is not sufficient detail to assess his instruments. Some *F*-statistics corresponding to p-values below .001 reject the null of weak instruments when $l = 10$; others fail to reject it.

In their study of income growth and the risk of property expropriation, AJR's first-stage t-statistics on their instrument, mortality rates among settlers when the country was first colonized, range from 2.0 to 4.0 across their specifications. Unfortunately, the lowest t-statistics, which would not lead us to reject the null hypothesis that the instrument is weak, arise when contemporary health conditions are added to the model. If health conditions today and mortality rates when a country was first colonized are correlated, omitting contemporary health conditions would bias 2SLS. We return to AJR's data below.

In their study of compulsory education, Angrist and Kreuger estimate the return to education using several identification strategies. In one, they just use quarter-of-birth dummies. In others they interact those dummies with dummies for year of birth and state. In the most extensive set of interactions, they use 178 instrumental variables. The F-statistics in some of these specifications indicate that the instruments are weak, especially in the cases of many instruments.

Yogo, AJR, and Angrist and Kreuger appear to have weak instrument problems. 2SLS estimates may be seriously biased in these cases, and inference based on the estimated variance of the 2SLS estimates and the normal distribution are likely to be invalid. What are we to do when our instruments are weak? Because the answer to this

[35] Currie and Yelowitz overcome an unusual difficulty. The CPS contains no data on kids' academic or housing outcomes. It doesn't allow the second stage of 2SLS. Decennial census data track such outcomes, but contains no data on public housing participation, so it doesn't allow the first stage of 2SLS. Both the CPS and the decennial census contain the untroublesome explanators and the instrument that Currie and Yelowitz use. Currie and Yelowitz follow Angrist and Krueger (1991) and construct second stage instruments in the decennial census data using the reduced-form parameter estimates from the CPS data.

question is more definitive when it comes to inference, we turn next to inference with weak instruments. We then turn to estimation.

Inference with Weak Instruments

The most common test in econometric modeling examines the null hypothesis that a particular parameter takes on one specific value: H_0: $\beta_1 = \beta_1^*$. Marcelo Moreira has developed a new two-sided test procedure for such hypotheses that is based on instrumental variables. In two papers, Moreira, Stock, and Donald Andrews have shown that this two-sided test procedure, called the two-sided conditional likelihood ratio (CLR) test is nearly optimal (in terms of its power) within a broad class of two-sided test procedures for IV models, unless one has good prior reason for thinking one instrument more salient than others (AMS (2005b) and Andrews and Stock (2005)). AMS have also developed a one-sided CLR test which they call CLR1 (AMS(2004)).[36,37]

The CLR test retains its superiority whether the instruments are strong or weak. AMS have also shown that that there does not exist an optimal one-sided test procedure, but that in simulations the one-sided CLR test usually performs about as well as the best of other one-sided tests, whether instruments are weak or strong (AMS (2005a)). AMS argue persuasively that The CLR test should be the test of choice in IV applications. However, longtime users of 2SLS may prefer to stick with standard tests when their instruments are strong, despite the optimality of the CLR test. If instruments are weak, and there is a single troublesome variable, The CLR test certainly does seem to be the right test procedure to use. [38] How best to conduct inference about the coefficients of a subset from among several troublesome variables when instruments are weak remains an open question[39].

How does The CLR test differ from an ordinary likelihood ratio test? In a standard Chi-square likelihood ratio test, the critical value for the test statistic is a fixed number that does not change with the data in hand; only degrees of freedom and the significance level matter. But standard likelihood ratio tests (and standard 2SLS Wald

[36] Frank Kleibergen independently developed a testing strategy closely akin to Moreira's (Kleibergen 2002)). Kleibergen has extended this approach to non-linear moment conditions, creating a conditional GMM framework for inference with weak instruments (Kleibergen 2005a,b). Andrews and Stock speculate that Kleibergen's tests has the good power properties of CLR even when disturbances are heteroskedastic or serially correlated.

[37] Economists sometimes want to test whether a particular variable is exogenous. Moreira (2005) shows that tests of exogeneity and tests of a variable's coefficient are intimately related. One ought not conduct both tests. If we suspect a variable is endogenous and we want to test a claim about its coefficient, we should treat the variable as endogenous, rather than testing its exogeneity and then estimating the coefficient accordingly.

[38] Andrews, Moreira, and Stock also offer heteroskedasticity-robust CLR tests and Autoregressive/ heteroskedasticity-robust CLR tests. These tests are also robust to omitted instruments in equation (2), which the plain CLR is not. A Stata command for implementing the two-sided CLR test can be downloaded from within Stata. The programs are at Marcelo Moreira's Harvard Web site.

[39] Kleibergen (2004) offers a strategy applicable when the instruments for untested coefficients' variables are strong and the instruments for tested coefficients' variables are weak. Dufour and Taamouti (2005a,b) also deal with this problem. I know of no packaged software for implementing these approaches.

tests) suffer size distortions when instruments are weak. Moreira overcomes the size distortions in ordinary tests by adjusting the critical values so that for given data, the critical values used yield a correct significance level - his critical values are "conditioned", not constant. Hence the name of Moreira's test. Rather than a fixed critical value for the likelihood ratio test, Moreira uses an aptly chosen "critical value function" to obtain a critical value that yields in practice the declared significance level for the test.[40] The standard Wald tests that we are accustomed to using in IV estimation can also be "conditioned" to correct their sizes, but the two-sided CLR test proves to be more powerful than those conditioned Wald tests, and about as powerful as a test can be absent prior information about which instruments are particularly good to rely on. Furthermore, the one-sided CLR test proves generally superior to the one-sided Wald test.[41]

Yogo uses the CLR test in studying the intertemporal elasticity of substitution, ψ. Yogo is not the first economist to estimate the elasticity of intertemporal substitution using lagged economic variables as instruments. For example, Hall also regresses the growth in consumption on the real interest rate to estimate ψ (Hall (1988)), and Lars Hansen and Kenneth Singleton estimate the reverse regression, with the real interest rate as the dependent variable and the growth in consumption as the explanator to obtain $1/\psi$ (Hall and Singleton (1983)). These two regression approaches have created a long-standing puzzle: Regressions of consumption growth on the interest rate tend to yield *small* 2SLS estimates of ψ, but the reverse 2SLS regressions imply *large* estimates of ψ. Consequently, the range of plausible estimates of ψ has been large. Yogo uses the latest IV techniques to resolve this puzzle.

In the reverse regression, consumption growth is the troublesome variable. Yogo finds that consumption growth is very hard to predict – the instruments are very weak in the reverse regression. 2SLS is quite biased in finite samples, and inferences based on such 2SLS estimates are invalid. In contrast, Yogo finds that interest rates are somewhat more predictable – the instruments are less weak in regressions of consumption growth on the interest rate. 2SLS is relatively unbiased in the direct regressions, but t-tests based upon such 2SLS estimates remain invalid. Yogo uses the CLR test because of its power and its robustness to weak instruments. Yogo rejects the null hypothesis that $\psi = 1$. The robustness of The CLR test to weak instruments makes Yogo's test result more credible than the findings of earlier analysts.

Kleibergen (2002) and Staiger and Stock (1996) have analyzed the quarter-of-birth and earnings data of Angrist and Kreuger using estimators robust to weak instruments. Those studies find that the 2SLS estimators used by Angrist and Kreuger sometimes provided confidence intervals that were much too narrow.

Estimation with Weak Instruments

[40] When the number of instruments is large relative to the number of observations, The CLR test (and others) becomes size distorted, though it still retains its good power properties relative to other test procedures.

[41] While Moreira's two-sided CLR test is about as good as we can do, the one sided CLR test is sometimes dominated by other conditioned tests.

Econometric theorists agree that 2SLS is a poor estimation strategy when instruments are weak. Point estimates and confidence intervals based on 2SLS are likely to be misleading. The point estimates might suffer considerable bias and the estimated confidence intervals are likely to be too narrow.

The best strategy for constructing a confidence interval for the coefficient of a lone troublesome variable is to construct it from Moreira's two-sided CLR hypothesis testing procedure: build a (1-α) confidence interval as the set of coefficient values that would not be rejected in The CLR test at the α level of significance.[42] Because Moreira's two sided CLR test's power is about the highest among two-sided tests, the resulting confidence interval is about as narrow as a (1-α) IV-based confidence interval could be. How to build valid confidence intervals when there are multiple troublesome explanators remains an open question.

Yogo reports CLR-based confidence intervals for the intertemporal elasticity of substitution. Consonant with Yogo's finding that 2SLS is approximately unbiased when applied to the direct regressions of consumption growth on the interest rate, the CLR-based confidence intervals contain the direct regression 2SLS estimate of ψ in each of Yogo's eleven countries. The biased character of the reverse regression 2SLS estimates is exposed sharply in that the estimates of ψ implicit in the reverse regression 2SLS estimates are excluded from the CLR-based confidence intervals for ψ. The long-standing puzzle was due to 2SLS being applied with weak instruments.

Several well-known estimation procedures have proven to have poor qualities as IV estimators when instruments are weak. Both 2SLS and limited information maximum likelihood estimation (LIML) can perform poorly when instruments are weak. LIML's chief problem is that it far too often yields wildly wrong parameter estimates when instruments are weak; the problem stems from LIML's lack of finite moments. 2SLS simply ought not be used when instruments might be weak.

In 1977, Wayne Fuller proposed estimators that modified LIML to obtain finite moments (Fuller (1977)). Fuller's estimators differed from one another by a parameter a, $a > 0$. Two of Fuller's estimators have become particularly popular, those with a =1 or a = 4. When a = 1, Fuller's estimator is approximately unbiased. When a = 4, Fuller's estimator is biased, but its mean square error is less than when a = 1.[43] Both of these Fuller estimators have proven to perform reasonably well when used for point estimation, even when instruments are weak.[44] Theorists are increasingly endorsing Fuller's estimators as better choices than 2SLS when seeking point estimates, especially when

[42] A Stata command that builds confidence intervals from the two-sided CLR test can be downloaded from within Stata; the algorithm is from Mikusheva (2005). Mikusheva's program is at Marcelo Moreira's Harvard Web site. If the disturbances in equation (1) are heteroskedastic or serially correlated, then heteroskedasticity and serial correlation robust versions of the CLR test should be used to build confidence intervals.

[43] Hahn, Hausman, and Kuersteiner (2003).

[44] The Fuller estimators are available within Stata's ivreg2 command.

instruments are weak.[45] However, some econometricians (Anderson, Kunitomo, and Matsushita (2005), Angrist and Kreuger (1991)) support the limited information maximum likelihood estimator (LIML) because its median is approximately equal to the coefficient of interest. (Fuller's estimators are bias and mean-square-error corrected versions of LIML.) When a model is exactly identified, LIML and 2SLS are equal.

AJR's 2SLS estimates of the coefficient on the risk of expropriation in their model of countries' incomes range from 0.55 to 1.20, with the smaller values arising when current health conditions are included in the model. Fuller estimates of that same coefficient range from 0.49 to 1.11 across that same span of specifications.

Using the Weak to Uncover the Strong

Like the CLR test, an Anderson-Rubin statistic (Anderson and Rubin (1949)) can be used to test robustly the hypothesis that the troublesome variable does not matter. In general, the CLR test dominates the Anderson-Rubin test for this purpose.[46] The Anderson and Rubin statistic can also provide a robust test of over-identifying restrictions. Unlike other tests of over-identifying restriction, the Anderson-Rubin test is robust to weak instruments.[47]

When, at the significance level α, this Anderson-Rubin test rejects the hypothesis that none of the over-identifying variables belong in equation (1), the CLR software reports an empty Anderson-Rubin (1-α) confidence interval for the troublesome variable's coefficient. For example, in AJR's data, in a regression in which the dependent variable is the log of real 1995 income, and the explanators are their measure of expropriation risk and a 1994 index of malaria for the country, using as instruments early settler mortality plus dummy variables indicating the country's continent results in an empty Anderson-Rubin confidence interval when $\alpha = .05$. If we believe the early mortality rate is a valid instrument, the conditional Anderson-Rubin test reveals that one or more of the continent dummies belong in the income equation.

When AJR were writing, little was known about how to conduct inference and estimation in the face of weak instruments. With hindsight, we know, as noted earlier, that some of AJR's estimations suffered from weak instruments. Here we find that recent advances in technique provide a sounder basis for the conclusions reached in AJR. AJR's data on income growth and expropriation risk in former colonies provides a fine opportunity for using the Anderson-Rubin test of over-identifying restrictions. In these data, a weak instrument with arguably good validity enables us to resolve whether a

[45] For example, Andrews and Stock (2005) and Hahn, Hausman and Kuersteiner (2003). Hansen, Hausman, and Newey show that using Fuller's estimator and Bekker's estimated standard errors largely shields Fuller's estimator from biases that can arise when there are many instruments (HHN (2005)., Bekker (1994)).

[46] In the case of an exactly identified equation, the Anderson-Rubin test is equivalent to the CLR test. It is when the equation is over-identified that they differ.

[47] AJR test the exogeneity of early settler mortality, conditional on one or another other instrument being valid, but they do not use the AR test.

much stronger instrument of more questionable validity is, in fact valid. Moreover, in these data, there is a need for a stronger instrument, as we shall see.

With expropriation risk, continent dummies, and the 1994 malaria index in AJR's income equation, a standard 2SLS t-test of a zero coefficient on the expropriation risk variable in AJR's income equation yields a *p*-value of 0.249; on that basis we fail to reject the null hypothesis of no effect of expropriation risk on income. The CLR test, which was not available to AJR, yields a *p*-value of 0.021 for that same hypothesis, so based on this more powerful test, we do reject the null hypothesis of no effect.[48]

Unfortunately, the result of the CLR test when continent dummies appear in AJR's income equation is less simple than it might at first seem. The estimated 95% confidence interval for the 2SLS estimator in that specification is [-0.97, 3.68], which is probably too narrow because of the downward bias in the estimated 2SLS standard error. The estimated 95% confidence interval based on The CLR test is the union of two disjoint sets: $[-\infty,-0.34]$ and $[0.44,+\infty]$. In this instance, The CLR test rejects small effects of expropriation risk on income, but rejects neither large negative effects nor large positive effects. This result is disquieting to those of us accustomed to convex confidence intervals. How does it come about, and how might we avoid it?

First, how can disjoint confidence intervals occur in IV estimation? The problem is rooted in the appearance of the covariance between the instrument and the troublesome variable in the denominator of the 2SLS estimator. Begin with an extreme case in which the population covariance between the instrument and the troublesome variable is, in fact, zero, in a bivariate model with a single troublesome explanator. With no instrument correlated with the troublesome explanator, the equation is under-identified. In this case, a CLR-based confidence interval will tend to be $[-\infty, +\infty]$ – we can't identify any one slope value as more likely than any other. As we move only slightly away from a zero covariance between the instrument and the troublesome variable, it sometimes happens that large β_1 values can't be excluded, because they, like the small denominator of the 2SLS estimator, give rise to estimates of β_1 that are large in magnitude. In some of these cases of weak instruments, there is a range of values that can be rejected. The AJR data give rise to such a case.[49]

How might we overcome such an outcome? The AJR data allow for one successful strategy. Adding a stronger instrument to the model will sometimes narrow the CLR confidence intervals. Can we use the weak early settler mortality rate to uncover a stronger additional instrument? Yes, we can. The proportion of the population who are of European descent in the country in 1900 is another variable in the AJR data set. Might this variable serve as an additional, potentially stronger, instrument for the risk of

[48] Chernozhukov and Hansen (2005) also note that AJR's instrument is weak. They propose an alternative hypothesis testing procedure robust to weak instruments that relies on the reduced form equations.

[49] A strong prior about the sign of the coefficient being estimated can resolve some of the uncertainty associated with a disjoint confidence region. If *a priori* we exclude negative values for β_1, for example, it becomes irrelevant that the confidence region contains large negative values for β_1, as well as large positive values.

expropriation? Maybe yes, and maybe no. Glaeser, et al. note that early settlers might have brought with them human capital that spurred growth (Glaeser, et al. (2004)). If this were the case, the proportion of the population of European descent in 1900 might belong in the model as an explanator that reflects the country's early level of human capital. In this case, that variable would not be available as an instrument. But perhaps the proportion of early settlers of European descent in 1900 is not needed as a variable in the 1995 income equation. In that case, the 1900 population proportion of European descent would be available as an additional instrument.

Can we use the early settler mortality rate instrument, which plausibly identifies the income equation once current health conditions and early human capital are controlled for, to test the validity of the 1900 proportion of the population of European descent as an instrument that might prove stronger than early settler mortality? Can we safely exclude the 1900 European population proportion from the income equation, which would make it available for use as an instrument? We can. The Anderson-Rubin test fails to reject the null hypothesis that the proportion of the population of European descent in 1900 variable is not a variable in the 1995 income equation.[50]

Using both the early settler mortality rate and the proportion of the population of European descent in 1900, the CLR test yields a 95% confidence interval of [.37, 1.32]. The *p*-value for the null hypothesis of no effect of expropriation risk on income is .0002 according to the CLR test. Does using both instruments overcome the weak instruments problem? No. The first-stage *F*-statistic for these two instruments is 14.92. The two instruments together are much stronger than the early mortality rate alone, but they still risk considerable size distortion if used with 2SLS. The first stage *F*-statistic is, however, large enough to reject the claim that 2SLS's finite sample bias is more than 10% that of OLS. The 2SLS estimate of the effect of expropriation risk on income is 0.60; the Fuller estimates are 0.62 when a=1 and 0.56 when a=4.

The Anderson-Rubin test of over-identifying restrictions proves a useful tool. Not only can it assess the validity of multiple weak instruments, but when there is a weak instrument that is likely valid, the test can enable us to check the validity of stronger potential instruments whose validity is in doubt.

Weak and Invalid Instruments

As with 2SLS's finite sample biases when instruments are valid, 2SLS's biases when the instruments are invalid are exacerbated by a low population R^2 in equation (2). Hahn and Hausman (2005) shows that under the simplifying assumptions made for equation (3), if instruments are invalid, the asymptotic bias in 2SLS is

[50] AJR performed a related robustness check for the validity of the early settler mortality rate instrument. In one specification, they included the 1975 proportion of a country's population that was of European descent. The risk of expropriation variable's coefficient remained significant in that specification. Note that the validity of the early settler mortality rate instrument does not resolve questions raised by Glaeser, et al. about whether the risk of expropriation is a good measure of government institutions, per se.

$$\frac{Cov(\hat{Y}_2, \varepsilon)}{\tilde{R}^2 Var(Y_2)}.$$

The appearance of $\tilde{R}^2$ in the numerator magnifies the effect of any covariance between the instrument and the disturbance. If $\tilde{R}^2$ is very small, as can happen with weak instruments, 2SLS can be seriously biased by even a small correlation between the instrument and the disturbance. Consequently, weak instruments require particular care to establishing the validity of instruments – "almost" valid may not do.

This observation has implications for an argument sometimes made when lagged variable values are used as instruments. The argument is implicit above in "perhaps we are confident in AJRY's study that the levels of democracy at long lags are uncorrelated with the disturbances, and our sole worry is whether the levels at shorter lags are actually uncorrelated with the disturbances." Analysts sometimes use longer lags of potential instruments on the supposition that the longer lags reduce the possible correlation between the instrument and the disturbances in equation (1). However, if, when a longer lagged variable is made the instrument, the covariance of the instrument with the disturbances of equation (1) doesn't fall relatively more than $\tilde{R}^2$, then using a longer lagged instrument doesn't decrease 2SLS's bias; it increases it. Because more distant lags are more likely to be weakly correlated with the troublesome variable, using distant lags increases the prospect that any source of invalidity, even small ones, in the instrument threatens to undermine 2SLS. Consequently, the case made for the validity of multiply lagged variable values as instruments must be especially strong for IV results to be credible.[51]

4. Interpreting IV estimates

Interpreting instrumental variables results can require some care. The consequences when such care is not taken can be ugly.

When Hoxby reconciles her finding of small and statistically insignificant effects of class size on test scores with findings of larger effects in an education experiment, she notes: "One might attribute some of the difference in the results to the necessarily transitory nature of population variation (from the teachers' not students' point of view)" (Hoxby (2000), p. 1281). Unpacking this observation brings into sharp relief a potential pitfall of IV estimation: if economic agents are heterogeneous in how they respond to a troublesome variable's value, IV estimation may tell us about an atypical group's behavior.

[51] In a related vein, when many lags of numerous variables are used as instruments, the degree of over-identification can grow large relative to the sample size. Having many instruments increases the finite sample biases of both the coefficient estimates and their estimated standard errors of 2SLS. Hansen, Hausman, and Newey (2005) provide guidance for better estimation when there are many instruments.

Suppose that class enrollments have both "permanent" and transitory components, so that total enrollment $X = X^P + X^T$, where X^P is the permanent enrollment (or enrollment expected by the teacher) and X^T is transitory enrollment (or the surprise in enrollment). Assume the two components of enrollment are statistically independent of one another. Further suppose, contrary to Hoxby, that teachers respond quite differently to the two kinds of enrollment. For example, perhaps when permanent enrollment changes, teachers adjust to the changed circumstance in some optimal fashion (because they expect that investments in changing will yield returns for some years to come), but maybe when transitory enrollment changes occur, the teachers do not adapt. In this scenario, the relationship between test scores, Y, and enrollments might be

$$Y = \beta_0 + \beta_P X^P + \beta_T X^T + \varepsilon,$$

with β_p distinctly different from β_p. (Hoxby argues that these two effects are not much different from one another, so responses to enrollment are not, in fact, very heterogeneous. It is instructive, however, to consider the heterogeneous case, as Hoxby cautions us to.)

We could envision an effect on performance of total enrollment, X, above and beyond the effects of X^P and X^T, in which case we could write

$$Y = \beta_0 + \beta_1 X + \beta_P X^P + \beta_T X^T + u. \qquad (6)$$

It is instructive to consider the case in which X, X^P and X^T are all contemporaneously uncorrelated with u. Unfortunately, equation (6)'s slopes are not identified because the explanators are perfectly collinear. To identify the equation, let's suppose we know that β_1 equals zero.

In practice, an econometrician will ordinarily observe total enrollment, X, not its components, X^P and X^T. Consequently, the econometrician will, like Hoxby, estimate

$$Y = \beta_0 + \beta_1 X + \upsilon,$$

where $\upsilon = \beta_P X^P + \beta_T X^T + u$. The OLS estimator of β_1 converges in probability to

$$\text{plim}(\beta_1^{OLS}) = \beta_P \frac{\sigma_P^2}{\sigma_P^2 + \sigma_T^2} + \beta_T \frac{\sigma_T^2}{\sigma_P^2 + \sigma_T^2}$$

where the σ's refer to the variances of X^P and X^T, respectively. [52] Thus, we find that when X is contemporaneously uncorrelated with μ, OLS applied to a Hoxby-style regression

[52] The omitted variables X^P and X^T bias the OLS estimate of β_1 (which equals 0 by assumption). The omitted variables bias will be

$$\text{plim}(\beta_1^{OLS}) - 0 = \beta_P \hat{\gamma}_P + \beta_T \hat{\gamma}_T,$$

consistently estimates a weighted average of β_P and β_T. If $\beta_P = \beta_T$, there is no problem with the OLS estimates. If the two coefficients are unequal, but the variance of X^P is much larger than that of X^T across the schools and years in our population, then OLS applied to X yields a slightly biased estimator for β_P.

What happens if we follow Hoxby and use X^T as an instrument for X in the regression of test scores on total enrollment?[53] The instrument is uncorrelated with both $\beta_P X^P$ and u in the disturbance, but it is correlated with $\beta_T X^T$. Consequently, the probability limit of this IV estimator of β_1 is β_T.[54]

Thus, the linear IV estimator of the effect of total enrollment on class size is equal to the effect of transitory changes in enrollment, and is unrelated to the effect of permanent changes in enrollment.

If β_p and β_T differ, OLS and IV will estimate different effects of enrollments on test scores – and the IV estimator may not estimate the effect of interest to policy makers. As James Heckman, Sergio Urzua, and Edward Vytlacil (HUV) write, "in a heterogeneous response model, there is no guarantee that IV is any closer to the parameter of interest than OLS" (HUV 2004, p. 2).

A rewritten equation (6) helps generalize from Hoxby's model. Consider

$$Y_i = \beta_0 + \left[\beta_P \pi_i + \beta_T (1 - \pi_i)\right] X_i + \nu_i = \beta_0 + \beta_{1i} X_i + \nu_i \qquad (6')$$

where $\pi_i = X_i^P / X_i$. Equation (6') expresses Hoxby's model as a random coefficients model. The mean of the β_i in the population is called the "average partial effect of X in the population." Other weighted averages of β_i are called "local average partial effects".[55] In general, when agents' responses to a troublesome variable are heterogeneous, IV estimation can yield a different average of the realized coefficients than OLS would – IV estimation may consistently estimate a local average partial effect, and not the average

---

where $\hat{\gamma}_P$ is the coefficient obtained by regressing the omitted variable X^P on the included variable, X, and $\hat{\gamma}_T$ is the coefficient obtained by regressing the omitted variable X^T on the included variable, X. These regressions of X's components on X are regressions of the components on a mis-measured version of themselves (X). The true coefficients are one, but attenuation bias will reduce each from one:

$$\hat{\gamma}_P = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_T^2} \qquad \text{and} \qquad \hat{\gamma}_T = \frac{\sigma_T^2}{\sigma_P^2 + \sigma_T^2},$$

[53] Hoxby actually uses an estimate of X^T, but for simplicity, I assume she uses X^T itself,.

[54]

$$\text{plim}(\beta_1^{IV}) = \beta_1 + \frac{p\lim\left(\frac{1}{n}\sum x^T \beta_T x^T\right)}{p\lim\left(\frac{1}{n}\sum x^T x\right)} = \frac{\beta_T\, p\lim\left(\frac{1}{n}\sum x^T x^T\right)}{p\lim\left(\frac{1}{n}\sum x^T (x^P + x^T\right)} = \beta_T .$$

[55] See Wooldridge (2002).

partial effect in the population.[56] When agents' responses are heterogeneous, care must be given to interpreting whose average response IV estimation consistently estimates.[57]

Sometimes a local average partial effect is not the effect we seek. Other times it is the effect we seek. For example, consider Hoxby's work on peer effects. Recall that Hoxby noted that surprises in enrollments are transitory for teachers, but not for students. In particular, students in a cohort with a surprisingly large number of girls in their class one year are likely to have a large number of girls in their class the next year, too. Hoxby's IV estimates of peer effects consistently estimate the effect of persistently having more girls or fewer girls across several years. Policymakers considering all girl and all boy classes might be interested in precisely the local average partial effect uncovered in Hoxby's work.

This same concern arises in Levitt's work. Levitt's instrumental variable estimates probably tell us much about the effects on crime rates of policies that reduce prison populations in a fashion much akin to the reductions made when overcrowding lawsuits are filed. Levitt's estimates are much less likely to be informative about the effects of releasing a markedly different subset of prisoners – in the extreme, for example, a program that replaces incarceration with probation for all sexual predators is apt to affect crime rates quite differently than Levitt's estimates would suggest. In brief, IV estimates are most reliably about policies of interest to us when the quasi-experimental treatment that the instrument defines is most like the real-world policy we envision.

Angrist and Krueger are quite conscious of the heterogeneity issue in their study of compulsory education. Their sample is the men in the 5% Public Use Micro Sample from the 1980 U.S. Census, but they assert that the effect of education on earnings estimated by their earnings equation is the effect of compulsory schooling on men's earnings – not the more general effect of education on earnings in the whole population of men. This is because "Two-stage least squares (TSLS) estimates are used in which the source of identification is variation in education that results solely from differences in season of birth – which, in turn, results from the effect of compulsory schooling laws." (p. 981)

The estimation of local partial effects has arisen most often in the context of program evaluations. In that context, local partial effects are called "local average treatment effects (LATE)" (Imbens and Angrist (1994)).

---

[56] When the troublesome explanator is correlated with the random coefficients, as in the Hoxby example, OLS is unlikely to consistently estimate the average partial effect in the population. In the Hoxby example, OLS does not converge in probability to the mean of β_{1i}, which is $\beta_P \bar{\pi} + \beta_T (1 - \bar{\pi})$, where $\bar{\pi} = E(\pi_i)$.

[57] In section 2, one strategy for assessing instruments validity was to ask whether several instruments give the same estimate. In the context of heterogeneous responses, two instruments uncorrelated with the *v* in equation 6' might identify different local average effects. If we believe the two instruments weight observations markedly differently from one another, getting similar coefficient estimate from them used separately would indicate that the variance in responses is not large.

In KLK's analysis of the Moving to Opportunity (MTO) experiment's effects on the youths' arrests, KLK estimate expected arrests (equation (3)) using the youth's status as experimental or control as an instrument. Section 2 argues that experimental status is a valid instrument in this example. What are the consequences of heterogeneous responses of youths to moving in this case? If the effect of using a voucher on youths' expected number of arrests varies across youth's, the 2SLS estimation measures the average effect of a voucher on those youths whose household uses a voucher in the experiment, but would not have used a voucher in the absence of being an experimental household. The average effect for these youths need not be the average effect across all youths; it is in this sense that the estimated local average effect is "local". If our policy interest is in the average effect of requiring all households to use vouchers, the estimated effect may not be not the effect of interest to us. Indeed, if our policy plan is to spread voucher use to any group different from the group who were induced to use vouchers in the MTO experiment, the estimates in KLK may not be informative. However, to the extent that we believe the experiment induces voucher use by the same sort of people as will be induced to use vouchers by our planned policies, the KLK local average treatment effects do tell us what we want to know.

The local average treatment effect model exploits the fact that both the troublesome variable and the instrument are dummy variables to clarify what the instrumental variables estimator actually estimates consistently. The Hoxby example similarly simplified the structure of the problem by assuming just two parameters of interest, β_P and β_T, that give a known structure to the random coefficients. Joshua Angrist, Guido Imbens, and Katherine Graddy (AIG) show in the context of supply and demand that some such structure is needed – as is clarity about what it is that we want to estimate – to make sense of instrumental variables estimators when responses of agents are heterogeneous (AIG (2000).

Heckman and Vytlacil have extensively analyzed estimating heterogeneous responses are in the contexts of program evaluation and returns to education (Heckman and Vytlacil (2005)); key papers have been co-authored with Pedro Carneiro and Sergio Urzua (CHV (2005) and HUV (2004)). These papers emphasize that when individuals respond heterogeneously, it is essential that we be clear what average or marginal effect we want to estimate. Heckman and Vytlacil explore when IV estimation can – and cannot! - identify the effects of interest and examine in detail the case in which the troublesome variable and the heterogeneous response are interdependent (Heckman and Vytlacil (2005)). A test for such interdependence is in HUV (2004).[58]

If responses are heterogeneous, the interpretation of rejections in classic over-identification becomes ambiguous. Over-identification tests reject the validity of over-identifying restrictions when the several instruments yield significantly different estimates of a troublesome variable's parameter. But significantly different IV estimates might result from instruments that identify different local average effects. This is an important consequence of a theorem first proved by Imbens and Angrist (1994).

[58] The Web site http://jenni.uchicago.edu/underiv/ contains documentation and a file with Fortran code for implementing the treatment effects estimation procedures of HUV (2004).

Consequently, rejections in over-identification tests are ambiguous: they might reflect invalid instruments or they might indicate heterogeneous responses.

This paper began with seven steps to take when using an instrumental variable. The lesson of this section is that two steps should precede using an instrumental variable:

i. If you anticipate heterogeneous responses among economic agents, ponder deeply whose responses and what kinds of responses are of economic interest. Whenever possible, model the responses to facilitate such judgments.

ii. Seek instruments that will expose the responses that are of economic interest.

In the words of HUV (2004, p.2): "In a model with essential heterogeneity, different instruments, valid for the homogeneous response model, identify different parameters. The right question to ask is "what parameter or combination of parameters is being identified by the instrument?", not "what is the efficient combination of instruments for a fixed parameter?", the traditional question addressed by econometricians."

5. Conclusion

IV estimation can be a powerful tool for overcoming biases that arise in OLS when a troublesome explanator is contemporaneously correlated with the disturbances. However, a promising instrument does not remove the need for carefulness in empirical analysis. Establishing an instrument's validity and relevance, coping with the possibility that an instrument is weak, and deciding whether the local effect estimated by IV estimation is the effect sought, all require imagination, diligence, and sophistication. The task is especially hard when instruments are weak, because weak instruments are particularly vulnerable to being cripplingly bad, but every IV analysis must worry whether the instrument selected is an appropriate one for estimating the effects of economic interest. The barriers to Archimedes moving the world were more daunting than the challenges facing IV estimation, but the comparison remains apt.

BIBLIOGRAPHY

Acemoglu, D., S. Johnson, and J. Robinson (2002), "Reversal of Fortune," *Quarterly Journal of Economics*,

Acemoglu, D., S. Johnson, and J. Robinson (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91:5, 1369-1401.

Acemoglu, D., and S. Johnson (2005), "Unbundling Institutions," *Journal of Political Economy*, Vol. 113, No.5, October, 949-995.

Acemoglu, D., S. Johnson, J. Robinson, and P. Yared (2005a), "From Education to Democracy," *American Economic Review*, forthcoming.

Acemoglu, D., S. Johnson, J. Robinson, and P. Yared (2005b), "Income and Democracy," NBER working paper #11205, March.

Anderson, T. W., and H. Rubin (1949), "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics*, 21, 570-582.

Anderson, T.W., N. Kunitomo, and Y. Matsushita (2005), "A New Light from Old Wisdoms: Alternative Estimation Methods of Simultaneous Equations and Microeconometric Models," CIRJE working paper F-321, University of Tokyo, February.

Andrews, D. W. K., and J. H. Stock (2005), "Inference with Weak Instruments," Cowles Foundation Discussion Paper, No. 1530, July, Yale University.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2005a), "Optimal Invariant Similar Tests for Instrumental Variables Regression with Weak Instruments," Cowles Foundation Discussion Paper, No. 1476, Yale University.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2005b), "Optimal Two-sided Invariant Similar Tests for Instrumental Variables Regression," June, mimeo.

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic* Review, Vol. 80, No. 3, June, 313-336.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000), "Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499-527.

Angrsit, J., and A. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, November ; reprinted in *Labor Economics*, O. Ashenfelter, ed., Edward Elgar Publishing Ltd, 1994; *Economic Demography*, T.P. Schultz, ed., Edward Elgar Publishing Limited, 1997; and the *Worth Series in Outstanding Contributions to Labor Economics*, 1999.

Angrist, Joshua D., Alan B. Krueger (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15:4, pp. 69-85.

Angrist, J., and A. Krueger (1991), "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, 328-337.

Angrist, J., and V. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement," *Quarterly Journal of Economics*, May; reprinted in *The Economics of Schooling and School Quality*, R. Hanushek, ed., Edward Elgar, 2002.

Angrist, J., and W. Evans (1998), "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, June; reprinted in *Recent Developments in Labor Economics*, J. Addison, ed., 2006.

Arellano, M., and S. R. Bond (1991), "Some Specification Tests for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277-298.

Arellano, M., L. P. Hansen, and E. Sentana (1999), "Underidentification?" mimeo, July.

Barro, R. J. (1999), "The Determinants of Democracy," *Journal of Political Economy*, 107, S158-S183.

Bekker, P. A. (1994), "Alternative Approximations to the Distribution of Instrumental Variables Estimators," *Econometrica*, 63, 657-681.

Baum, C. F., M. E. Schaffer, and S. Stillman (2003). "Instrumental Variables and GMM: Estimation and testing", Boston College Working Paper Series, No. 545.

Campbell, J. Y., and L. M. Viciera (2002), *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, Clarendon Lectures in Economics, New York, Oxford University Press.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2005), "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," mimeo, May.

Chernozhukov, V, and C. Hansen (2005), "The Reduced Form: A Simple Approach to Inference with Weak Instruments," working paper, University of Chicago.

Cruz, L. M., and M. Moreira (2005), "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws," *Journal of Human Resources*, 40(2), 393-410.

Currie, J., and A. Yelowitz (2000), "Are Public Housing Projects Good for Kids?" *Journal of Public Economics*, 75:1, 89-124.

Dufour, J., and M. Taamouti (2005a), "Further results on projection-based inference in IV regressions with weak, collinear or missing instruments," *Journal of Econometrics*, forthcoming.

Dufour, J., and M. Taamouti (2005b), "Projection-Based Statistical Inference In Linear structural Models With Possibly Weak Instruments," *Econometrica*, Vol. 73, No. 4, July, 1351–1365.

Epstein, L. G., and S. E. Zin (1989), "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica*, 57:4, 937-969.

Fuller, W. A. (1977), "Some Properties of a Modification of the Limited Information Maximum Likelihood Estimator," *Econometrica*, 45, 939-954.

Glaeser, E. L., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer (2004, "Do Institutions Cause Growth?" NBER Working Paper #10568.

Hahn, J. and J. Hausman (2002): "A New Specification Test for the Validity Instrumental Variables," *Econometrica* 70, 163 – 189.

Hahn, J., and J. Hausman (2005), "Instrumental Variable Estimation with Valid and Invalid Instruments," July, Cambridge, MA, mimeo.

Hahn, J., J. Hausman, and G. Kuersteiner (2004), "Estimation with Weak Instruments: Accuracy of Higher Order Bias and MSE Approximations", *Econometrics Journal*, 7, 272-306.

Hall, R. E. (1988), "Intertemporal Substitution in Consumption," *Journal of Political Economy*, 96:2, 339-357.

Hansen, C., J. Hausman, and W. Newey (2005), "Estimation with Many Instrumental Variables," July, Cambridge MA, mimeo.

Hansen, L. P., and K. J. Singleton (1983), "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, 91:2, 249-265.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hausman, J. (1983), "Specification and Estimation of Simultaneous Equation Models," Chapter 7 in *Handbook of Econometrics, vol. 1*, Z. Griliches and M. D. Intriligator, eds., 391-448.

Heckman, J. J., S. Urzua, and E. Vytlacil (2004), "Understanding Instrumental Variables In Models With Essential Heterogeneity," mimeo, December.

Heckman, J. J., and E. Vytlacil (2005), "Structural Equations, Treatment Effects And Econometric Policy Evaluation," *Econometrica*, .

Hoxby, C. M. (2000), "The Effect of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115, November, 1239-1285.

Hoxby, C. M. (2002), "The Power of Peers: How Does the Makeup of a Classroom Influence Achievement," *Education Next*, Summer, 2002.

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

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781-1803.

Kleibergen, F. (2004), "Testing Subsets of Structural Parameters in the Instrumental Variables Regression Model," *Review of Economics and Statistics*, 86, 418-423.

Kleibergen, F. (2005a), "Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics," working paper, Department of Economics, Brown University.

Kleibergen, F. (2005b), "Testing Parameters in GMM without Assuming That They Are Identified," *Econometrica*, 73, forthcoming.

Kling, J. R., J. Ludwig, and L. F. Katz (2005), "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment," *Quarterly Journal of Economics*, 120:1, February, 87-130.

Leamer, E. E. (1978) *Specification Searches: Ad Hoc Inference with Non-Experimental Data*, New York, NY, John Wiley and Sons.

Levitt, S. D. (1996), "The effect of Prison Population Size on Crime rates: Evidence from Prison Overcrowding Litigation." *Quarterly Journal of Economics*, 111:2, May, 319-351.

Levitt, S. D. (1997), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 87:4, June, 270-290.

Levitt, S. D. (2002), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review*, 92:4, June, 1244-1250.

McCrary, (2004), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment," *American Economic Review*, 92:4, June, 1236-1243.

Miller, A. J. (1990), *Subset Selection in Regression*, New York: Chapman-Hall.

Mikusheva, A. (2005), "An Algorithm for Constructing Confidence Intervals Robust to Weak Instruments," mimeo, Cambridge, MA, October.

Moreira, M. J. (2003), "A Conditional Likelihood Test for Structural Models," *Econometrica*, 71:4, 1027-1048.

Moreira, M. (20005), "Tests With Correct Size When Instruments Can Be Arbitrarily Weak," mimeo, Cambridge, MA, July.

Murray, M. P. (2004), *Econometrics: A Modern Introduction*, Boston, Addison-Wesley.

Murray, M. P. (1983), "Subsidized and Unsubsidized Housing Starts, 1961-1977" *Review of Economics and Statistics*, 1961-1977," 65:4, November, 590-597.

Nelson, C., and Startz, (1990a), "Some Further results on the Exact Small Sample Properties of the Instrumental Variables Estimator," *Econometrica*, Vol. 58, No. 4, 967-976.

Nelson, C., and Startz, (1990b), "The Distribution of the Instrumental Variables Estimator and Its f-Ratio When the instrument Is a Poor One,* *Journal of Business*, 1990, vol. 63, no. 1, pt. 2, S125-S140.

Phillips, P. C. B. (1983), "Exact Small Sample theory in the Simultaneous Equations Model," Chapter 7 in *Handbook of Econometrics*, vol. 1, Z. Griliches and M. D. Intriligator, eds., 449-516.

Rothenberg, T. J. (1983), "Asymptotic Properties of Some Estimators In Structural Models," in *Studies in Econometrics, Time Series, and Multivariate Statistics*, in honor of T. W. Anderson. Edited by S. Karlin, T. Amemiya, and L. A. Goodman, Academic Press, 1983.

Rothenberg, T.J. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics, "Chapter 15 in *Handbook of Econometrics*, vol. 2, Z. Griliches and M. D. Intriligator, eds., 881-935.

Sargan, J. D. (1958), "The Estimation of Economic relationships with Instrumental Variables," *Econometrica*, 26:3, July, 393-415.

Staiger, D., and J. H. Stock (1997), "Instrumental Variables regressions with Weak Instruments," *Econometrica*, May, 65:3, 557-586.

Stock, J. H., and M. Watson (2003), *Introduction to Econometrics*, Addison-Wesley, Boston, MA.

Stock, J. H., and M. Yogo (2005), "Testing for Weak Instruments in IV Regression," in *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg*, Cambridge University Press.

Stock, J. H., J. H. Wright, and M. Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20:4, 518-529.

Van der Klaauw, W. (1996), "A Regression-Discontinuity Analysis of the Effect of College Aid offers on College Enrollment," manuscript, New York University.

Woodford, M. D. (2003), *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton, Princeton University Press.

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MIT Press, 2002.

Wooldridge, Jeffrey M. (2000), *Introductory Economics: A Modern Approach*, South-Western Publishing, 2000.

Yogo, M. (2004), "Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak," *Review of Economics and Statistics*, 86:3, August, 797-810.