ESTIMATING AND TESTING MODELS WITH MANY TREATMENT LEVELS AND LIMITED INSTRUMENTS

Lance Lochner and Enrico Moretti*

Abstract—Empirical researchers interested in the causal effect of the endogenous regressor often use instrumental variables. When few valid instruments are available, they typically estimate restricted specifications that impose uniform per unit treatment effects, even when these effects are likely to vary. We show that in these cases, ordinary least squares and instrumental variables estimators identify different weighted averages of all per unit effects, so the traditional Hausman test is uninformative about endogeneity. We develop a new exogeneity test that works even when the true model cannot be estimated using IV methods as long as a single valid instrument is available. We revisit three recent empirical examples to demonstrate the practical value of our test.

I. Introduction

MANY recent empirical papers seek to estimate causal relationships using instrumental variables (IV), including two-stage least squares (2SLS) estimators, when concerns about causality arise. A model frequently estimated in practice has the following form:

$$y_i = s_i\beta^L + x_i'\gamma^L + \nu_i, \tag{1}$$

where y_i is the outcome of individual i, x_i is a $k \times 1$ vector of exogenous covariates (including an intercept), and s_i is the potentially endogenous regressor. For example, the variable s_i might reflect different treatment levels of a government training program or different dosage levels for a new drug treatment. In our empirical examples and much of our discussion, s_i reflects years of completed schooling.

Conclusions about exogeneity of s_i and consistency of the ordinary least squares (OLS) estimator are typically based on a comparison of OLS and IV estimates of β^L. When a standard Hausman test (Hausman, 1978) indicates a significant difference between OLS and IV estimates, it is common to conclude that endogeneity of s_i plays an important confounding role in OLS.

Yet in many economics applications, the true relationship between y_i and s_i is unlikely to be linear. In particular, suppose that the endogenous regressor $s_i \in \{0, 1, 2, 3, ..., S\}$ is discrete and the true model has the form

$$y_i = \sum_{j=1}^{S} D_{ij}\beta_j + x_i'\gamma + \varepsilon_i, \tag{2}$$

Received for publication May 3, 2011. Revision accepted for publication March 26, 2014. Editor: Alberto Abadie.

* Lochner: University of Western Ontario and NBER; Moretti: University of California–Berkeley, NBER, CEPR, and IZA.

We thank Josh Angrist, David Card, Pedro Carneiro, Jim Heckman, Guido Imbens, the editor, two anonymous referees, and seminar participants at the 2008 UM/MSU/UWO Summer Labor Conference, UCSD, and Stanford for their suggestions. We also thank Matias Cattaneo and Javier Cano Urbina for their excellent research assistance and comments, as well as Martijn van Hasselt and Youngki Shin for their many comments and suggestions.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00475.

where $D_{ij} = 1[s_i \geq j]$ reflects a dummy variable equal to 1 if $s_i \geq j$ and 0 otherwise, $E(\varepsilon_i) = 0$, and $E(\varepsilon_i x_i) = 0$. When s_i reflects years of schooling, the β_j represent grade-specific effects of moving from $j - 1$ to j years of schooling.

The difference between the models in equations (1) and (2) is that the former assumes a uniform per unit or marginal effect of s_i across all levels of s_i while the latter does not. For example, in the classic case of the return to education, the model in equation (1) assumes that the effect of an extra year of elementary school is identical to the effect of the last years of high school and college, while the model in equation (2) allows for sheepskin effects and other nonlinearities that are likely to arise in practice.

While variable per unit treatment effects are likely to be important in many applications, relatively few studies have focused on their practical implications when instrumental variables may be needed.[1] The difficulty in estimating a specification like equation (2) when endogeneity concerns arise is that there may be many β_j parameters to estimate, while researchers typically have very few valid instruments. In theory, a single continuous instrument may be sufficient for identification. In practice, there is often insufficient variation in the instrument to precisely estimate all per unit effects. Discrete-valued instruments are also common in the literature. As a consequence, empirical studies commonly estimate models like equation (1) even when there is no theoretical reason to do so, and in some cases there is prima facie evidence of important nonlinearities between y_i and s_i.

We demonstrate that when the per unit effects of changes in s_i vary over the range of s_i as in equation (2) but the estimated model assumes that all per unit effects are the same as in equation (1), OLS and IV methods estimate different weighted averages of all per unit effects. Building on this insight, we develop a new exogeneity test that requires only a single (even binary) instrument and is useful when per unit treatment effects vary across treatment levels.

We stress that our results do not apply to all nonlinear models, only to the specific case described in equations (1) and (2). In particular, we assume (a) a single finite-valued discrete endogenous regressor, (b) exogenous regressors are additively separable and enter the equation linearly; and (c) all coefficients (including per unit treatment effects) are homogeneous in the population. While these assumptions are strong, they are common in the applied microeconomics literature.

We are not the first to point out that estimates from a misspecified linear model (i.e., constant marginal or per unit treatment effects) yield weighted averages of each

[1] Angrist, Graddy, and Imbens (2000), Lochner and Moretti (2001), and Mogstad and Wiswall (2010) are notable exceptions.

The Review of Economics and Statistics, May 2015, 97(2): 387–397

doi:10.1162/REST_a_00475

marginal/per unit effect. Yitzhaki (1996) derives these weights in the context of OLS, while Angrist and Imbens (1995) and Heckman, Urzua, and Vytlacil (2006) derive weights for IV estimators in the presence of both variable multivalued treatment effects and parameter heterogeneity. Angrist and Imbens (1995) show conditions under which 2SLS estimates a local average treatment effect (LATE).[2] In a very general setting, Heckman et al. (2006) discuss ordered and unordered choice models with unobserved heterogeneity and nonlinearity, developing weights for treatment effects using general instruments. Heckman and Vytlacil (2005) emphasize that in the presence of parameter heterogeneity, there is no single effect of the regressor on an outcome, and different estimation strategies provide estimates of different parameters of interest or different average effects. While many studies focus on parameter heterogeneity across individuals with a uniform marginal effect over values of s_i (i.e., y_i is linear in s_i), we consider the opposite case, assuming a nonlinear relationship between y_i and s_i that is the same for all individuals.[3] Our setting is a special case of that used by Heckman et al. (2006); however, our emphasis on varying per unit treatment effects, and the endogeneity test is novel.

We begin by showing that inappropriately assuming model (1) when per unit treatment effects vary across treatment levels will generally yield different OLS and IV/2SLS estimates even in the absence of endogeneity, since these estimators can be written as weighted averages of causal responses to each marginal change in the regressor, where the sets of weights differ for the estimators.[4] An appealing feature of our setting is that the weights have an intuitive interpretation, are functions of observable quantities, and can be easily estimated under very general assumptions. Therefore, it is possible to directly compare the OLS and IV weights.

This insight leads to our main contribution: a new exogeneity test that can be used to determine the consistency of the OLS estimator for equation (2). Before describing our test, first note that the standard Hausman test is of limited applicability in this context. Since OLS and IV/2SLS identify different weighted averages of all per unit effects, the Hausman test applied to equation (1) is uninformative about the endogeneity of the regressor when per unit treatment effects vary across treatment levels. It may reject equality of OLS and IV/2SLS estimates even when the regressor is exogenous, and it may fail to reject equality when the regressor is endogenous. Alternatively, in order to implement the Hausman test for equation (2), one would need to estimate all β_j parameters using IV methods. In practice, this is often impossible when there are many treatment levels, since researchers often have access to only a few valid instruments with limited variation. Rarely would researchers have instruments capable of identifying, for example, twenty different grade-specific β_j parameters associated with all potential schooling levels.

The test that we propose can be thought of as a generalization of the standard Hausman test and is informative about the consistency of OLS estimates for all β_j effects in equation (2). Our test reweights OLS estimates of the β_js from equation (2) using estimated IV/2SLS weights and compares this with the corresponding IV/2SLS estimator of β^L in equation (1). Under fairly general conditions, our test can be implemented even when only a single valid (binary) instrument is available.[5]

Our proposed test has both strengths and weaknesses. The fact that our test requires only a single instrument should make it attractive to empirical researchers. In many contexts, researchers can easily use OLS to estimate models like equation (2) (e.g., regressing log wages on a set of twenty schooling dummies), yet they often have very few valid instruments with limited variation at their disposal. A researcher can use our test to establish whether the OLS estimates are consistent without having to estimate the more general equation (2) using IV/2SLS. If our test fails to reject exogeneity, researchers can have some confidence in their OLS estimates. However, if our test rejects, it does not help in estimating the true model. Our test therefore offers only a partial solution to the problem of estimating multiple per unit treatment effects with limited instruments.

Three additional limitations are worth highlighting. First, it is important to note that we test whether the weighted average of all OLS $\hat{\beta}_j$ asymptotic biases equals 0. Therefore, our test has no power against the possibility that some OLS $\hat{\beta}_j$ estimates are asymptotically biased upward and others downward in such a way as to exactly cancel each other when averaged using the IV/2SLS weights. Still, rejection of the null implies that OLS estimates are inconsistent. Furthermore, we discuss conditions under which all $\hat{\beta}_j$ asymptotic biases would be of the same sign, in which case our test is equivalent to testing whether all OLS $\hat{\beta}_j$ estimates are consistent. In many applications, economic theory can be informative about the likely sign of any biases. For example, in the case of returns to schooling, most models of investment in human capital predict that OLS estimates of $\hat{\beta}_j$ are all asymptotically upward biased.

[2] Intuitively, the LATE reflects the effect of a regressor on outcomes for individuals induced to change their behavior in response to a change in the value of the instrument.

[3] Studies focused on parameter heterogeneity include Imbens and Angrist (1994), Wooldridge (1997), Heckman and Vytlacil (1998, 1999, 2005), Card (1999), Kling (2001), Moffitt (2009), and Carneiro, Heckman, and Vytlacil (2010).

[4] Relative to the existing literature, our models are closer to those typically estimated in practice. Angrist and Imbens (1995) consider only discrete regressors that are indicators that place observations into mutually exclusive categories, and they interact their instrument (also assumed to be discrete) with each of these regressors to create a large set of effective instruments. The Heckman et al. (2006) discussion of instrumental variables estimation in ordered-choice models is left implicit on all covariates affecting the outcome variable.

[5] Lochner and Moretti (2001) and Mogstad and Wiswall (2010) suggest that comparing reweighted OLS estimates with IV/2SLS estimates may be a useful heuristic approach for assessing the importance of nonlinearities. In this paper, we develop a formal econometric test for exogeneity based on this insight. Our test differs conceptually and practically from the omnibus specification tests developed by White (1981), which essentially compare different weighted generalized least squares estimators for a general nonlinear function.

Second, even if exogeneity cannot be rejected, researchers should exercise caution when conducting inference using OLS estimates of equation (2) when the instruments are not sufficiently strong. Like the Hausman test, our test does not have much power when instruments are weak. As Wong (1997) and Guggenberger (2010) demonstrate, this can cause size problems with inference in a two-stage approach where the Hausman test is used to determine exogeneity in a first stage and OLS estimates are used in a second stage when exogeneity cannot be rejected.[6] Monte Carlo simulations confirm that similar inference problems can arise when using our test with insufficiently strong instruments.

Third, our approach assumes that equation (2) reflects the true model. Misspecification due to, for example, nonseparabilities between s_i and x_i or due to individual-level parameter heterogeneity would likely invalidate our test, since this would alter the relationship between OLS and IV estimators in unaccounted-for ways.

In the last part of the paper, we demonstrate the practical usefulness of our test by reexamining three recent empirical papers in which estimated 2SLS effects differ from OLS effects. In one example, our test suggests that schooling is exogenous for incarceration among white men. As we discuss, this is empirically useful since it lends credibility to OLS estimates that suggest a highly nonlinear relationship between educational attainment and the probability of imprisonment. In contrast, our test strongly rejects exogeneity of schooling for incarceration among black men, while the standard Hausman test does not. In this case, the endogeneity of schooling is obscured when nonlinearities between schooling and imprisonment are ignored. Our other examples produce greater concordance between the standard Hausman test and our exogeneity test, although for different reasons.

The rest of the paper is organized as follows. In section II, we show conditions under which OLS, IV, and 2SLS estimates of β^L in equation (1) can be written as weighted averages of the true underlying β_j parameters in the more general model given by equation (2). For expositional purposes, we will refer to s_i as years of schooling, so the β_j reflect grade-specific marginal or per unit effects. Section III develops an exogeneity test that can be used to determine consistency of the OLS estimator for equation (2). Section IV presents the results from three previous empirical examples, and section V concludes.

II. Estimating Weighted Average Per Unit Treatment Effects

In this section, we consider IV/2SLS and OLS estimators when equation (1) is estimated, but the true model is described by equation (2). We show conditions under which these estimators converge to a weighted average of each grade-specific β_j effect and discuss the weights. We assume throughout our analysis that all observations are independent across $i = 1, \ldots, N$ individuals and that standard conditions for the weak law of large numbers and central limit theorems apply.[7]

A. IV Estimation with a Single Instrument

We first consider IV estimation with a single instrument, discussing OLS as a special case. We study the case where the potentially endogenous variable s_i is discrete.[8] Throughout the paper, we assume ε_i is independent across individuals with $E(\varepsilon_i) = 0$, x_i is distributed with density $F_x(\cdot)$, and $E(\varepsilon_i x_i) = 0$. The following decomposition is also useful: $s_i = x_i'\delta_s + \eta_i$, where $\delta_s = [E(x_i x_i')]^{-1}E(x_i s_i)$ by construction and $E(x_i\eta_i) = 0$.

The following IV assumption is standard.

Assumption 1. The instrument is uncorrelated with the error in the outcome equation, $E(\varepsilon_i z_i) = 0$, and correlated with s_i after linearly controlling for x_i, $E(\eta_i z_i) \neq 0$.

Let $M_x = I - x(x'x)^{-1}x'$ and $\tilde{s} = M_x s$ for any variable s. (We drop the i subscripts when we refer to the vector or matrix version of a variable that vertically stacks all individual-specific values.) With a single instrument, 2SLS estimation of equation (1) is equivalent to the following IV estimator:

$$\begin{aligned}\hat{\beta}^L_{IV} &= (z'M_x s)^{-1}z'M_x y \\ &= (\tilde{z}'\tilde{s})^{-1}\tilde{z}'\left(\sum_{j=1}^{S} D_j\beta_j\right) + (\tilde{z}'\tilde{s})^{-1}\tilde{z}'\varepsilon \\ &= \sum_{j=1}^{S}\hat{\omega}^{IV}_j\beta_j + (\tilde{z}'s)^{-1}\tilde{z}'\varepsilon,\end{aligned}$$

where $\hat{\omega}^{IV}_j = (\tilde{z}'\tilde{s})^{-1}\tilde{z}'D_j = \left(\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_i D_{ij}\right) \Big/ \left(\frac{1}{N}\sum_{i=1}^{N}\tilde{z}_i\tilde{s}_i\right)$. Since $\sum_{j=1}^{S} D_{ij} = s_i$, these $\hat{\omega}^{IV}_j$ sum to 1 over $j = 1, \ldots, S$. We refer to them as weights even though they may be negative for some j.[9]

One helpful assumption is monotonicity in the effects of the instrument on s_i. Although monotonicity is not necessary

[6] Specifically, Wong (1997) and Guggenberger (2010) provide simulation evidence in a linear regression model like equation (1) for the null rejection probability of a simple hypothesis test conditional on a standard Hausman pretest for exogeneity not rejecting. Their findings indicate that when regressor endogeneity is small, the null rejection probability of the hypothesis test may be substantially higher than the nominal size if the instruments are not sufficiently strong.

[7] For example, assume all random variables are independent and have finite first, second, and third moments. Finite third moments enable application of central limit theorems based on independent but not necessarily identically distributed random variables (e.g., Liapounov).

[8] While we study the case of a discrete endogenous regressor, OLS and IV estimators will also yield different weighted averages of marginal effects when the regressor is continuous. The insights of Yithzaki (1996) might be used to develop weights and a related test specifically designed for the continuous regressor case.

[9] When they cannot be shown to be nonnegative, we use "weights" with quotation marks to distinguish them from cases when they are known to be proper weights that are both nonnegative and sum to 1.

for deriving and estimating "weights," it does help to ensure that they are nonnegative and simplifies their interpretation. When s_i reflects years of schooling, monotonicity implies that the instrument either causes everyone to weakly increase or causes everyone to weakly decrease their schooling. Without loss of generality, we assume that s_i is weakly increasing in z_i. Define $s_i(\vartheta)$ to be the value of s_i for individual i when $z_i = \vartheta$.

Assumption 2 (monotonicity)**.** The instrument does not decrease s_i: $Pr[s_i(\vartheta) < s_i(\vartheta')] = 0$, for all $\vartheta > \vartheta'$.

To facilitate our analysis of $\hat{\beta}^L_{IV}$, it is useful to decompose $z_i = x_i'\delta_z + \zeta_i$ where $\delta_z = [E(x_i x_i')]^{-1}E(x_i z_i)$ and $E(x_i\zeta_i) = 0$.

Proposition 1. *If assumption 1 holds, then* $\hat{\beta}^L_{IV} \xrightarrow{p} \sum_{j=1}^{S} \omega_j^{IV}\beta_j$*, where*

$$\omega_j^{IV} = \frac{Pr(s_i \geq j)E(\zeta_i|s_i \geq j)}{\sum_{k=1}^{S}[Pr(s_i \geq k)E(\zeta_i|s_i \geq k)]} \tag{3}$$

sum to unity over all $j = 1, \ldots, S$*. Furthermore, if* $E(z_i|x_i) = x_i'\delta_z$ *and assumption 2 (monotonicity) holds, then the weights are nonnegative and can be written as*

$$\omega_j^{IV} = \frac{E\{\mathrm{Cov}(z_i, D_{ij}|x_i)\}}{\sum_{k=1}^{S} E\{\mathrm{Cov}(z_i, D_{ik}|x_i)\}} \geq 0. \tag{4}$$

Proof. See online appendix A.

This result shows that estimating the misspecified linear-in-schooling model using IV yields a consistent estimate of a weighted average of all grade-specific β_j effects. The weights on all grade-specific effects are straightforward to estimate. From a 2SLS regression of D_{ij} on s_i and x_i using z_i as an instrument for s_i, the coefficient estimate on s_i equals $\hat{\omega}_j^{IV}$.

When the instrument affects all persons in the same direction and its expectation conditional on x_i is linear (e.g., x's are mutually exclusive and exhaustive categorical indicator variables), the weights are nonnegative and depend on the strength of the covariance between the instrument and each schooling transition indicator conditional on other covariates. In general, different instruments yield estimates of different "weighted averages," even if the instruments are all valid.

With assumption 1, $E(z_i|x_i) = x_i'\delta_z$, and $E(\varepsilon_i|x_i) = 0$, it is straightforward to show that the IV estimator converges to a weighted average of all conditional (on x_i) IV estimators, where the weights are proportional to the covariance between the instrument and schooling conditional on x_i:

$$\hat{\beta}^L_{IV} \xrightarrow{p} \int \beta_{IV}(\phi)h(\phi)dF_x(\phi),$$

where $\beta_{IV}(\phi) = \frac{\mathrm{Cov}(z_i,y_i|x_i=\phi)}{\mathrm{Cov}(z_i,s_i|x_i=\phi)}$ is the population analog of the IV estimator conditional on $x_i = \phi$ and $h(\phi) = \frac{\mathrm{Cov}(z_i,s_i|x_i=\phi)}{\int \mathrm{Cov}(z_i,s_i|x_i=a)dF(a)}$ is a weighting function that integrates to 1 for all x_i (with $h(\cdot) \geq 0$ under assumption 2). Notice that $\beta_{IV}(\phi) = \sum_{j=1}^{S} \beta_j\omega_j^{IV}(\phi)$, where $\omega_j^{IV}(\phi) = \frac{\mathrm{Cov}(z_i,D_{ij}|x_i=\phi)}{\mathrm{Cov}(z_i,s_i|x_i=\phi)}$ are x-specific IV "weights" for each grade-specific effect, β_j. Each x-specific IV estimator is simply a weighted average of the grade-specific β_j effects, where the weights are proportional to the covariance between the instrument and D_{ij} conditional on x_i. Some rearranging shows that the IV weights from equations (3) or (4) can be rewritten as $\omega_j^{IV} = \int \omega_j^{IV}(\phi)h(\phi)dF_x(\phi)$.[10]

These results complement the IV/2SLS analyses of Angrist and Imbens (1995) and Heckman et al. (2006), who also consider parameter heterogeneity along with variable per unit treatment effects. In order to ease interpretation in the presence of parameter heterogeneity, Angrist and Imbens (1995) make strong assumptions about the additional x_i covariates and how they enter in estimation. Specifically, they assume that the x_i regressors are indicator variables that place individuals into mutually exclusive categories and that the instrumental variable (also assumed to be discrete) is interacted with all of these additional covariates. By contrast, Heckman et al. (2006) consider a very general setting for ordered and unordered choice models; however, their discussion of IV estimation for these models implicitly conditions on all covariates x_i (deriving IV weights analogous to $\omega_j^{IV}(\phi)$ in our setting). Results in this section could therefore be derived as a special case of their analysis. While our analysis ignores heterogeneity in the grade-specific effects, it considers estimation under common assumptions about covariates and the way they typically enter during estimation. We are not focused on finding an "economic interpretation" for the IV estimator, since the weights we consider can easily be estimated. Instead, we are interested in empirically comparing the OLS and IV weights and deriving a test for whether the different weights can explain differences between the two estimators when per unit treatment effects are incorrectly assumed to be uniform (i.e., linearity between y_i and s_i).

Since OLS is a special case of IV estimation, in the absence of endogeneity, the OLS estimator for the linear-in-s_i model in equation (1) also converges to a weighted average of the grade-specific effects, β_j, where the weights are nonnegative and sum to 1.

Corollary 1. *If* $E(\varepsilon_i s_i) = 0$ *then*

$$\hat{\beta}^L_{OLS} \xrightarrow{p} \sum_{j=1}^{S} \omega_j^{OLS}\beta_j, \tag{5}$$

[10] In online appendix A, we further show that with a binary instrument, the $\omega_j^{IV}(\cdot)$ weights can be more easily interpreted along the lines of the LATE analysis of Angrist and Imbens (1995).

where the

$$\omega_j^{OLS} = \frac{Pr(s_i \geq j)E(\eta_i|s_i \geq j)}{\sum_{k=1}^{S} Pr(s_i \geq k)E(\eta_i|s_i \geq k)} \geq 0 \tag{6}$$

sum to unity over all $j = 1, \ldots, S$.

Proof. This result largely follows from proposition 1 replacing z_i with s_i. Online appendix A shows that the OLS weights are always nonnegative.

The empirical counterpart to the OLS weights, $\hat{\omega}_j^{OLS} \xrightarrow{p} \omega_j^{OLS}$, is simply the coefficient estimate on s_i in an OLS regression of D_{ij} on s_i and x_i. Therefore, only data on x_i and s_i are needed to construct consistent estimates of the asymptotic weights. Of course, the weights implied by OLS estimation will not generally equal the weights implied by IV estimation.[11] In section IV, we graph estimated OLS and IV weights in a few different empirical applications.

Researchers often estimate models like equation (1) rather than the more general equation (2) because they are limited in the instrumental variables at their disposal. Yet even in the absence of endogeneity and individual-level parameter heterogeneity, there is no reason to expect OLS and IV estimators to be equal for a misspecified linear-in-s_i model that assumes uniform per unit treatment effects. As a result, standard Hausman tests applied to the mis-specified linear-in-s_i model may reject the null hypothesis of "exogenous s" due simply to variable per unit treatment effects. Below, we develop a chi-square test for whether OLS estimation of equation (2) yields consistent estimates of the underlying β_j parameters (i.e., whether $E(\varepsilon_i|s_i) = 0$) even when only a single valid instrumental variable is available. However, we first generalize our key results to the case of many instruments.

B. 2SLS Estimation with Multiple Instruments

We now generalize the results to the case where we have I distinct instruments for schooling, $z_i = (z_{i1} \ldots z_{iI})'$, but the researcher still estimates the linear-in-schooling model (1). Let $s_i = x_i'\theta_x + z_i'\theta_z + \xi_i$, with $\hat{\theta}_x$ and $\hat{\theta}_z$ reflecting the corresponding OLS estimates of θ_x and θ_z. Further define the predicted value of schooling conditional on x and z: $\hat{s}_i = x_i'\hat{\theta}_x + z_i'\hat{\theta}_z$. Then 2SLS estimation of equation (1) yields

$$\hat{\beta}_{2SLS}^L = (\hat{s}'M_x\hat{s})^{-1}\hat{s}'M_xy = \sum_{j=1}^{S} \hat{\omega}_j\beta_j + (\hat{s}'M_x\hat{s})^{-1}\hat{s}'M_x\varepsilon, \tag{7}$$

where the "weights" $\hat{\omega}_j = (\hat{s}'M_x\hat{s})^{-1}\hat{s}'M_xD_j = (\hat{\theta}_z'z'M_xz\hat{\theta}_z)^{-1}\hat{\theta}_z'z'M_xD_j$ reflect consistent estimates of ω_j from 2SLS estimation of

$$D_{ij} = s_i\omega_j + x_i'\alpha_j + \psi_{ij}, \qquad \forall j \in \{1, \ldots, S\}. \tag{8}$$

We assume that assumption 1 holds for all $z_{i\ell}$ instruments and that we have sufficient variation in z_i conditional on x_i for identification. Let $\zeta_i = (\zeta_{i1}, \ldots, \zeta_{iI})'$ be the $I \times 1$ vector collecting all $\zeta_{i\ell} = z_{i\ell} - x_i'\delta_{z\ell}$, where $\delta_{z\ell} = [E(x_ix_i')]^{-1}E(x_iz_{i\ell})$ was introduced above in the single-instrument case.[12]

Assumption 3. The covariance matrix for z_i after partialling out x_i, $E(\zeta_i\zeta_i')$, is full rank.

As with the single-instrument IV estimator, we can show that the 2SLS estimator for β^L in equation (1) converges in probability to a "weighted" average of all grade-specific effects. Letting $\omega_{j\ell}^{IV}$ reflect the grade j "weight" from the single-instrument IV estimator using $z_{i\ell}$ as the instrument as defined by equation (3), the 2SLS estimator "weight" on any β_j is a weighted average of each of these single-instrument IV estimator "weights":

Proposition 2. *Under assumptions 1 and 3, $\hat{\beta}_{2SLS}^L \xrightarrow{p} \sum_{j=1}^{S} \omega_j\beta_j$, where $\omega_j = \sum_{\ell=1}^{I} \Omega_\ell\omega_{j\ell}^{IV}$ sum to unity over all $j = 1, \ldots, S$ and*

$$\Omega_\ell = \frac{\theta_{z\ell}\sum_{k=1}^{S} Pr(s_i \geq k)E(\zeta_{i\ell}|s_i \geq k)}{\sum_{m=1}^{I}\theta_{zm}\sum_{k=1}^{S} Pr(s_i \geq k)E(\zeta_{im}|s_i \geq k)} \tag{9}$$

sum to unity over all $\ell = 1, \ldots, I$. Furthermore, if each instrument satisfies assumption 2 and $E(z_{i\ell}|x_i) = x_i\delta_{z\ell}$, then all $\omega_{j\ell}^{IV}$, Ω_ℓ, and ω_j are nonnegative.

Proof. See online appendix A.

Not surprisingly, one can also show that the 2SLS estimator converges in probability to a weighted average of the probability limits of all single-instrument IV estimators, where the weights are given by Ω_ℓ in equation (9).[13]

[11] For example, consider the case with no x regressors (except an intercept). It is straightforward to show that $\omega_{j+1}^{OLS} - \omega_j^{OLS} \propto (E(s_i) - j) \times Pr(s_i = j)$, which is positive for $j < E(s_i)$, 0 for $j = E(s_i)$, and negative when $j > E(s_i)$. This implies that OLS estimation of the linear specification places the most weight on grade-specific β_j effects near the mean schooling level. When schooling is uniformly distributed in the population, the weights decay symmetrically as one moves away from the mean in either direction. Contrast this with the IV weights in the case of a binary instrument $z_i \in \{0, 1\}$ satisfying the monotonicity assumption. In this case, IV places all the weight on schooling margins that are affected by the instrument, while the underlying distribution of schooling in the population is irrelevant.

[12] In the case of a single instrument, this analysis reduces to that for IV in section IIA with $\hat{\beta}_{2SLS}^L = \hat{\beta}_{IV}^L$ and $\hat{\omega}_j = \hat{\omega}_j^{IV}$ for all j.

[13] If we define $\beta_{IV,\ell}^L = plim\ \hat{\beta}_{IV,\ell}^L$ where $\hat{\beta}_{IV,\ell}^L$ is the single-instrument IV estimator using $z_{i\ell}$ as an instrument for s_i in estimating equation (1), then $\hat{\beta}_{2SLS}^L \xrightarrow{p} \sum_{\ell=1}^{I} \Omega_\ell\beta_{IV,\ell}^L$, where Ω_ℓ is defined by equation (9).

III. A Wald Test for Consistent OLS Estimation of All β_j's

When at least one valid instrumental variable is available, the analysis of section II suggests a practical test for whether OLS estimates of $B \equiv (\beta_1, \ldots, \beta_S)$ from equation (2), $\hat{B}$, are consistent.[14] We now develop a test that compares the 2SLS estimator from equation (1) with the weighted sum of the grade-specific OLS estimates of the β_js from equation (2), using the estimated 2SLS weights $\hat{\omega} \equiv (\hat{\omega}_1, \ldots, \hat{\omega}_S)'$. Intuitively, if $E(\varepsilon_i|s_i) = 0$, so the grade-specific OLS estimates are consistent, then the reweighted sum of these OLS estimates (using the 2SLS weights) should asymptotically equal the 2SLS estimator from equation (1), that is, $\hat{\beta}^L_{2SLS} - \hat{\omega}'\hat{B} \xrightarrow{p} 0$. This will not generally be true when $E(\varepsilon_i D_{ij}) \neq 0$ for any j.

Applying 2SLS to equation (8) yields estimates $\hat{\omega}_j$ and $\hat{\alpha}_j$ for all j. In order to derive our test statistic, we frame estimation of $\hat{B}$, $\hat{\beta}^L_{2SLS}$, and $\hat{\omega}$ as a stacked generalized method of moments (GMM) problem. This establishes joint normality of $(\hat{B}, \hat{\beta}^L_{2SLS}, \hat{\omega})$ and facilitates estimation of the covariance matrix for all of these estimators. From this, a straightforward application of the delta method yields the variance of $\hat{\beta}^L_{2SLS} - \hat{\omega}'\hat{B}$, which is used in developing a chi-square test statistic for the null hypothesis that $\hat{T} \equiv \hat{\beta}^L_{2SLS} - \hat{\omega}'\hat{B} \xrightarrow{p} 0$.

It is necessary to introduce some additional notation in order to define the test statistic. We first define the regressors for OLS estimation of equation (2), $X_{1i} = (D_i'\ x_i')$, and the regressors, $X_{2i} = (s_i\ x_i')$, and instruments, $Z_{2i} = (z_i'\ x_i')$, used in 2SLS estimation of equations (1) and (8). Denote the corresponding matrices for all individuals as X_1, X_2, and Z_2, respectively. Next, let $\Theta = (B'\ \gamma'\ \beta^L\ \gamma^{L'}\ \omega_1'\ \alpha_1'\ \ldots\ \omega_S'\ \alpha_S')'$ reflect the full set of parameters to be estimated. Finally, let $\hat{\Theta}$ denote the corresponding vector of parameter estimates, where $(B'\ \gamma')$ is estimated by OLS and $(\beta^L\ \gamma^{L'})$ and all $(\omega_j'\ \alpha_j')$ are estimated via 2SLS.

The variance of Θ can be consistently estimated from

$$\hat{V} = \hat{A}\hat{\Lambda}\hat{A}', \tag{10}$$

where

$$\hat{A} = \begin{pmatrix} [X_1'X_1]^{-1} & \mathbf{0} \\ \mathbf{0} & I_2 \otimes [\hat{X}_2'\hat{X}_2]^{-1}\hat{\Gamma}_2' \end{pmatrix}, \tag{11}$$

$\hat{\Gamma}_2 = (Z_2'Z_2)^{-1}Z_2'X_2$, $\hat{X}_2 = Z_2\hat{\Gamma}_2$, $\boldsymbol{I}_2$ is an $S+1$ dimension identity matrix, and $\mathbf{0}$ reflects conformable matrices of zeros.[15] Furthermore,

$$\hat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N} \times \begin{pmatrix} \hat{\varepsilon}_i^2(X_{1i}'X_{1i}) & \hat{\varepsilon}_i\hat{\nu}_i(X_{1i}'Z_{2i}) & \hat{\varepsilon}_i\hat{\Psi}_i' \otimes (X_{1i}'Z_{2i}) \\ \hat{\varepsilon}_i\hat{\nu}_i(Z_{2i}'X_{1i}) & \hat{\nu}_i^2(Z_{2i}'Z_{2i}) & \hat{\nu}_i\hat{\Psi}_i' \otimes (Z_{2i}'Z_{2i}) \\ \hat{\varepsilon}_i\hat{\Psi}_i \otimes (Z_{2i}'X_{1i}) & \hat{\nu}_i\hat{\Psi}_i \otimes (Z_{2i}'Z_{2i}) & \hat{\Psi}_i\hat{\Psi}_i' \otimes (Z_{2i}'Z_{2i}) \end{pmatrix}, \tag{12}$$

where $\hat{\varepsilon}_i = y_i - D_i'\hat{B} - x_i'\hat{\gamma}$, $\hat{\nu}_i = y_i - s_i\hat{\beta}^L_{2SLS} - x_i'\hat{\gamma}^L$, and $\hat{\Psi}_i = (\hat{\psi}_{1i}\ \hat{\psi}_{2i} \ldots \hat{\psi}_{Si})'$ with $\hat{\psi}_{ij} = D_{ij} - s_i\hat{\omega}_j - \hat{\alpha}_j'x_i$.

Finally, define $\hat{T} \equiv T(\hat{\Theta}) = \hat{\beta}^L_{2SLS} - \hat{\omega}'\hat{B}$, and let

$$\hat{G} \equiv \nabla\hat{T} = (-\hat{\omega}'\ 0_x'\ 1\ 0_x'\ (-\hat{\beta}_1\ 0_x')\ (-\hat{\beta}_2\ 0_x')\ \ldots\ (-\hat{\beta}_S\ 0_x'))$$

represent the $(2S+1+(S+2)K) \times 1$ jacobian vector for $T(\hat{\Theta})$ (where 0_x is a $K \times 1$ zero vector).

It is now possible to derive a chi-square test statistic:

Theorem 1. *Under assumptions 1 and 3, if $E(\varepsilon_i|s_i) = 0$, then*

$$W_N = N\left[\frac{(\hat{\beta}^L_{2SLS} - \hat{\omega}'\hat{B})^2}{\hat{G}\hat{V}\hat{G}'}\right] \xrightarrow{d} \chi^2(1). \tag{13}$$

Proof. See online appendix A.

It is important to note that $\hat{T} \xrightarrow{p} 0$ need not imply that $\hat{B} \xrightarrow{p} B$ for two reasons. First, this test cannot tell us anything about whether $\hat{\beta}_j \xrightarrow{p} \beta_j$ for some grade transition j if $\omega_j = 0$. The test provides information only about the effects of grade transitions that are affected by the instrument. Second, the $\hat{\beta}_j$ OLS estimates may be asymptotically biased upward for some j and downward for others. When $E(\varepsilon_i|s_i) \neq 0$, $\hat{B} \xrightarrow{p} B^* \equiv B + \{E(D_iD_i') - E(D_ix_i')[E(x_ix_i')]^{-1}E(x_iD_i')\}^{-1}E(D_i\varepsilon_i)$. Thus, $\hat{T} \xrightarrow{p} 0$ for any B^* satisfying $\omega'(B - B^*) = 0$, where $\omega \equiv (\omega_1, \ldots, \omega_S)'$. A test based on theorem 1 would have no power against these alternatives, although rejection of the null hypothesis would imply that $\hat{B}$ does not consistently estimate B.

Under reasonable conditions, W_N can serve as a valid test statistic for the null hypothesis that $\hat{B} \xrightarrow{p} B$. If $\omega_j > 0$ for all j (a testable assumption) and if $E(\varepsilon_i D_{ij}) = E(\varepsilon_i|s_i \geq j)\ Pr(s_i \geq j)$ were either nonnegative for all j or nonpositive for all j, then all $\hat{\beta}_j$ would be asymptotically biased in the same direction and $B^* \neq B \Leftrightarrow \omega'(B - B^*) \neq 0$. In this case, testing whether $\hat{T} \xrightarrow{p} 0$ would be equivalent to testing for consistency of $\hat{B}$.[16]

To better understand these conditions, consider a standard latent index ordered-choice model for schooling of the form

$$s_i^* = \mu(z_i, x_i) + v_i, \tag{14}$$

$$s_i = j \quad \text{if and only if } j \leq s_i^* < j+1. \tag{15}$$

Assume that all x regressors and instruments z are independent of both errors: $(\varepsilon_i, v_i) \perp\!\!\!\perp (z_i, x_i)$. It is straightforward to show that if $E(\varepsilon_i|v_i)$ is weakly monotonic in v_i, then $E(\varepsilon_i|s_i \geq j)$ will be either nonpositive or nonnegative for all j.[17] Mono-

[14] Formally, $\hat{B} = (D'M_xD)^{-1}D'M_xy$, where M_x and y are defined earlier and D reflects the stacked $N \times S$ matrix of $(D_{i1}, \ldots, D_{iS})$ for all individuals.

[15] See the proof of theorem 1 in online appendix A.

[16] In the case where some $\omega_j = 0$, the test would be equivalent to testing for consistency of all β_j with $\omega_j > 0$.

[17] Strictly speaking, weak monotonicity is required only over the range of v_i covered by $j - \mu(z_i, x_i)$ (i.e., for $v_i \in [1 - \mu(z_i, x_i), S - \mu(z_i, x_i)]$), so behavior in the tails of the distribution is irrelevant. See online appendix A for details.

tonicity of $E(\varepsilon_i|v_i)$ is trivially satisfied by all joint elliptical distributions (e.g., bivariate normal or t distributions), which produce linear conditional expectation functions.

In practice, one is likely to fail to reject the null hypothesis of $\hat{T} \xrightarrow{p} 0$ when $B^* \neq B$ only in cases where individuals with both high and low propensities for education (conditional on observable characteristics) have a higher (or lower) unobserved ε_i than individuals with an average propensity for schooling. In the case of an ordered-choice model, this would imply a U-shaped (or inverted U-shaped) relationship for $E(\varepsilon_i|v_i)$. In many economic contexts, these perverse cases seem unlikely.

We also note that if more than one valid instrument is available, then those instruments can be used in different combinations to perform separate tests. Because each 2SLS estimator (distinguished by the set of instruments used) converges to a different weighted average of the true B parameters (i.e., $\omega'_\Upsilon B$ where Υ denotes the set of instruments used), it is unlikely that one would reject the null of $\omega'_\Upsilon B = \omega'_\Upsilon B^*$ for all sets of instruments unless $B = B^*$.[18]

To demonstrate the extent to which varying per unit treatment effects can induce differences between OLS and IV estimates that our new exogeneity test can account for (while standard Hausman or Durbin-Wu-Hausman tests applied to equation (1) cannot), we perform a Monte Carlo simulation exercise based on Card's (1995) log earnings–schooling model. In this framework, varying per unit treatment effects are equivalent to a nonlinear relationship between log earnings and schooling. These results are discussed in detail in online appendix B; however, we note here that our test (see theorem 1) performs well in two important respects. First, the test has nearly identical performance to the standard Hausman test, applied to equation (1), when all grade-specific effects are the same. Thus, there is no cost to using our test rather than the more traditional Hausman test that assumes a linear relationship between log earnings and schooling. Second, our test has very similar properties regardless of the extent of nonlinearity between log earnings and schooling, rejecting equality of the reweighted OLS and IV estimates at noticeably higher rates for even small deviations from exogeneity as long as the instruments are sufficiently strong.

Of course, when the instruments are relatively weak, our test (like the standard Hausman test) has little power to detect endogeneity since the IV estimates tend to have large standard errors. In these cases, negligible amounts of endogeneity may be difficult to detect with our test. This can lead to poor size properties when conducting inference using OLS estimates of the β_j parameters as discussed by Wong (1997) and Guggenberger (2010), who study this issue in the context of linear models and use of the Hausman test to determine exogeneity. Monte Carlo results presented in online appendix B suggest caution when using OLS estimates for inference, even if our test fails to reject exogeneity, if the instruments are relatively weak. This is particularly true when the IV and reweighted OLS estimates are quite different but the IV estimates are very imprecise.

Another important limitation to keep in mind is that our test is valid only if equation (2) represents the true model. This model assumes that the regressors are additively separable and that the coefficients are the same for all individuals. In the case of nonseparability or individual heterogeneity in the model's coefficients, our model would be misspecified and our test invalid.

IV. Practical Use of Our Test and Three Empirical Examples

To demonstrate the practical value of our test, we reexamine three empirical papers on the effects of individual and maternal schooling that estimated 2SLS effects that differ nontrivially from their corresponding OLS estimates.[19] In all cases, the econometric specification assumed a linear relationship between the outcome of interest and educational attainment as in equation (1).[20] Of course, if the true relationship is nonlinear so grade-specific effects differ, then differences between OLS and 2SLS weights may explain at least some of the difference between the two estimates. For each of the three cases, we examine the extent to which reweighting the OLS estimates of the β_js helps reconcile the difference between the potentially misspecified OLS and 2SLS estimates that assume uniform grade-specific effects. We then test whether schooling is exogenous using both the standard Hausman test and our proposed test.

Results are reported in table 1.[21] Columns 1 and 2 reproduce OLS and 2SLS estimates using the same models and similar data used in the original papers. For example, the first row indicates that when the Lochner and Moretti (2004) data for white men are used, a regression of an indicator for incarceration on years of schooling and controls yields an OLS coefficient equal to $-.0010$, and a 2SLS coefficient equal to $-.0011$. The 2SLS estimates use as instrumental variables three indicators for different compulsory schooling ages. The difference between OLS and 2SLS is reported in column 3. The 2SLS estimate is about 10% larger than the

[18] Because these test statistics are not generally independent, the critical values for this type of joint testing procedure are likely to be quite complicated. We do not address this issue here.

[19] The instruments used in these examples have been employed in numerous studies examining a wide array of outcomes. See Lochner (2011).

[20] In two of the applications we consider (Lochner & Moretti, 2004; Currie & Moretti, 2003), the outcome variables are binary, and a linear probability model is assumed by the authors. Heteroskedasticity of errors does not pose any problems for our test; however, our assumption of separability between all regressors and measures of schooling is questionable in more general binary choice models for well-known reasons. We simply follow the specifications employed in the earlier studies, assuming the data are consistent with a linear probability model. This may not be unreasonable in these applications given the limited range of predicted outcome probabilities across values of the regressors. Assuming an index model based on equation (2), the density for the error may be (approximately) linear over the range of estimated index values.

[21] Details regarding samples and estimating specifications are reported in the bottom of table 1.

TABLE 1.—REPLICATION RESULTS AND APPLICATION OF WALD TEST FOR ENDOGENEITY

	$\hat{\beta}^L_{OLS}$	$\hat{\beta}^L_{2SLS}$	$\hat{\beta}^L_{OLS}-\hat{\beta}^L_{2SLS}$	$\sum_j \hat{\omega}_j\hat{\beta}_j$	General Wald Test Statistic [p-value]	Hausman Test Statistic [p-value]	DWH Test Statistic [p-value]
Lochner and Moretti (2004): Effect of years of schooling on imprisonment							
White males	−0.0010	−0.0011	−0.0002	−0.0012	0.0225	0.2021	0.1600
	(0.0000)	(0.0004)	(0.0004)	(0.0000)	[0.8808]	[0.6530]	[0.6858]
Black males	−0.0037	−0.0048	−0.0011	−0.0007	11.9441	0.9757	0.5154
	(0.0001)	(0.0012)	(0.0011)	(0.0002)	[0.0005]	[0.3233]	[0.4728]
Currie and Moretti (2003): Effect of maternal education on infant health							
Low birth weight	−0.0050	−0.0098	−0.0048	−0.0053	1.4376	1.7022	1.5566
	(0.0001)	(0.0038)	(0.0037)	(0.0002)	[0.2305]	[0.1920]	[0.2122]
Preterm birth	−0.0044	−0.0104	−0.0060	−0.0046	1.7639	2.0472	1.7749
	(0.0002)	(0.0044)	(0.0042)	(0.0002)	[0.1841]	[0.1525]	[0.1828]
Acemoglu and Angrist (2001): Private returns to schooling							
Annual earnings	0.0822	0.1442	0.0620	0.0832	5.7093	6.0028	6.0218
	(0.0003)	(0.0256)	(0.0253)	(0.0017)	[0.0169]	[0.0143]	[0.0141]

The first four columns report estimates for reported parameters with standard errors in parentheses. Columns for general Wald test, Hausman, and Durbin-Wu-Hausman (DWH) test report test statistics (p-values are in brackets) for the null hypothesis of exogeneity. General Wald test compares $\hat{\beta}^L_{2SLS}$ and $\sum_j \hat{\omega}_j\hat{\beta}_j$ as described in theorem 1, while the Hausman and DWH tests compare $\hat{\beta}^L_{2SLS}$ and $\hat{\beta}^{OLS}_L$. Specifications for Lochner and Moretti (2004) use men ages 20 to 60 from the 1960 to 1980 U.S. Censuses and include indicators for three-year age categories, year, state of birth, and state of residence. Specifications for blacks also include an indicator for whether the individual turned age 14 after 1957 and was born in the South. Specifications from Currie and Moretti (2003) use first-time white mothers ages 24 to 35 from Vital Statistics Natality records from 1970 to 1999 and include median county income, percent urban in county when the mother was 17, and indicators for ten-year birth cohorts, mother's age, and county-specific year of child's birth effects. Specifications for Acemoglu and Angrist (2001) results differ slightly from theirs, since we use only compulsory attendance indicators for instruments and do not estimate the social return to schooling. Specifications use white men 40 to 49 year old from the 1960 to 1980 U.S. Censuses and include indicators for Census year, year of birth, state of birth, and state of residence.

OLS estimate (in absolute value), even though most reasonable explanations for the endogeneity of schooling suggest that the OLS estimate should overstate the importance of schooling. The corresponding OLS and 2SLS estimates for blacks are −.0037 and −.0048, respectively.

There are several well-understood reasons that one might find a larger 2SLS estimate (relative to the OLS estimate), including the presence of measurement error and individual-level heterogeneity in the effects of schooling. It is also possible that nonlinearity in the incarceration-schooling relationship may play a role. This seems particularly relevant here given the pattern of OLS estimates for the grade-specific effects β_j reported in figures 1 and 2. If the assumption of uniform grade-specific effects were correct and these estimates were consistent, all of the estimated $\hat{\beta}_j$ should be the same. Instead, the estimated $\hat{\beta}_j$ suggest that the marginal effects of different grade transitions vary considerably across years of schooling. Unless there are much stronger biases for some grades than others, the figures suggest strong nonlinearities in the relationship between imprisonment and schooling, with the strongest effect for high school graduation (moving from grade 11 to 12).[22] Based on these findings, Lochner and Moretti (2004) suggest that high school graduation is an important margin for incarceration among men, but they are hesitant to draw strong conclusions from these OLS estimates due to concerns about endogeneity.

The lines in figures 1 and 2 report estimates of the OLS and 2SLS weights, as defined in section II. These weights are clearly very different for white men: the OLS weights are high between 12 and 16 years of schooling, while the

FIGURE 1.—EFFECTS OF SCHOOLING ON THE PROBABILITY OF INCARCERATION FOR WHITE MALES (OLS ESTIMATES AND WEIGHTS)

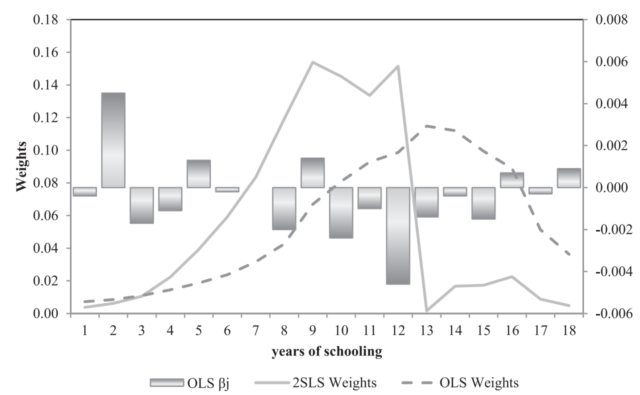
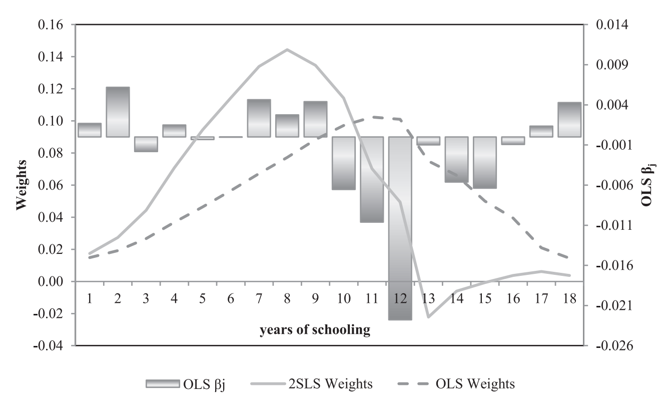


FIGURE 2.—EFFECTS OF SCHOOLING ON THE PROBABILITY OF INCARCERATION FOR BLACK MALES (OLS ESTIMATES AND WEIGHTS)



2SLS weights are highest at 12 years of schooling, implying that the effect of moving from 11 to 12 years of schooling figures prominently in the 2SLS estimates. This is not surprising since the instruments adopted (compulsory schooling laws) are most effective at shifting schooling levels just

[22] Standard errors for the β_j estimates are all less than 0.001 (0.003) for whites (blacks) except for the first two grade levels. Estimates along with their standard errors are reported in online appendix table C1. For comparison, we also report average marginal effects from analogous logit specifications in online appendix table C2. The pattern of effects is quite similar.

before or at high school graduation. For black men, the effect of compulsory schooling is strong at earlier grades, so that the weights are more shifted to the left. In column 4 of table 1, we reweight the estimated grade-specific effects ($\hat{\beta}_j$) using the 2SLS weights in figure 1.[23] For whites, the reweighted OLS estimates are -0.0012, larger than the 2SLS estimates. The reweighted OLS estimates are larger, because 2SLS puts more weight on the large β_j associated with moving from 11 to 12 years of schooling. For blacks, the reweighted OLS estimate is smaller, because the 2SLS weights are more shifted to the left and therefore put less weight on larger β_j.

The last three columns of table 1 are the most important, since they report on different tests for the exogeneity of schooling. Column 5 presents test statistics and associated p-values for our proposed test of exogeneity (see theorem 1), which is valid even when the effects of schooling differ across grades. Columns 6 and 7 present results from the standard Hausman test and the Durbin-Wu-Hausman test (applied to the linear-in-schooling specification), respectively, which are both incorrect when the grade-specific effects differ. For white men, our test fails to reject, which is quite important in practice, since it suggests that our OLS estimates of the β_j in figure 1 are consistent. Given a high first-stage F-statistic of 1,000.3 and the fact that the reweighted OLS estimate is very close to the 2SLS estimate (a difference of less than 10 percent), it seems reasonable to conclude from our OLS estimates of β_j that high school completion has the strongest effect on incarceration rates, while college attendance has much weaker effects.[24] This is extremely useful, since with our limited set of instruments, it is impossible to estimate all 20 β_j parameters using 2SLS. Indeed, 2SLS estimates from highly restricted two-parameter models that relax linearity in schooling are very imprecise. Fortunately, our test suggests that IV methods are not necessary in this case.

The case of incarceration for black men is different: our test strongly rejects the hypothesis that the reweighted OLS and 2SLS estimates are the same (p-value of .0005), while the standard Hausman test fails to reject. Reweighting the OLS estimates for the β_j parameters reveals that the OLS estimates are significantly biased toward 0, on average, since the reweighted OLS estimate is $-.0007$ compared to the 2SLS estimate of $-.0048$. In this case, we cannot draw any strong conclusions about the relative importance of different grades due to these biases. These findings empirically demonstrate that when grade-specific effects may differ, the standard Hausman test can fail to detect an endogeneity problem when one exists.

In the second panel, we turn to estimates of the effect of maternal schooling on infant health from Currie and Moretti (2003). The instrument in this case is an indicator for college proximity. (First-stage F-statistics for the instruments are 398.7.) In this case, the reweighted OLS estimates (column 4) are generally quite similar to the OLS estimates (column 1). Looking at figures 3 and 4, it is clear why: the OLS and 2SLS weights are nearly identical. Not surprisingly, our test and the standard Hausman test produce very similar test statistics and the same conclusions: exogeneity cannot be rejected for either child health outcome.

FIGURE 3.—EFFECTS OF MATERNAL SCHOOLING ON THE PROBABILITY OF LOW BIRTH WEIGHT (OLS ESTIMATES AND WEIGHTS)

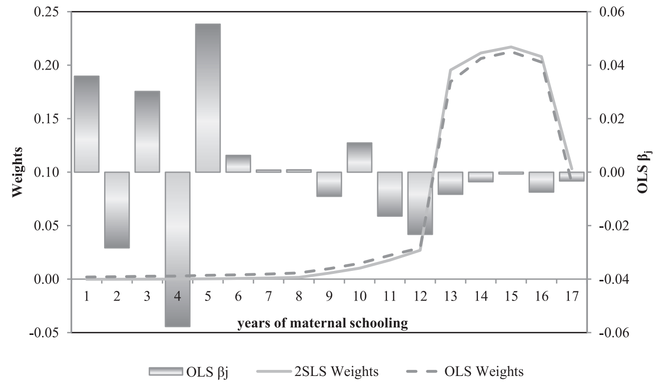
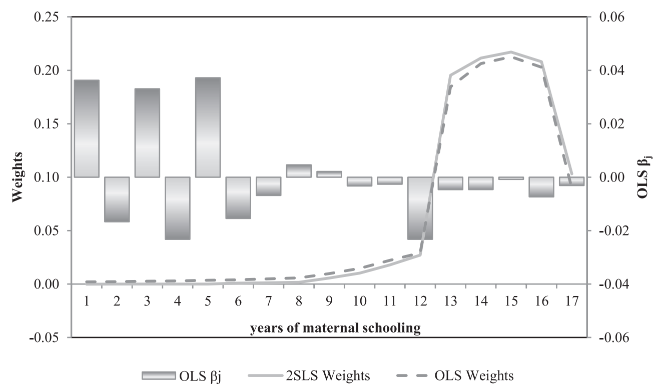


FIGURE 4.—EFFECTS OF MATERNAL SCHOOLING ON THE PROBABILITY OF PRETERM BIRTH (OLS ESTIMATES AND WEIGHTS)

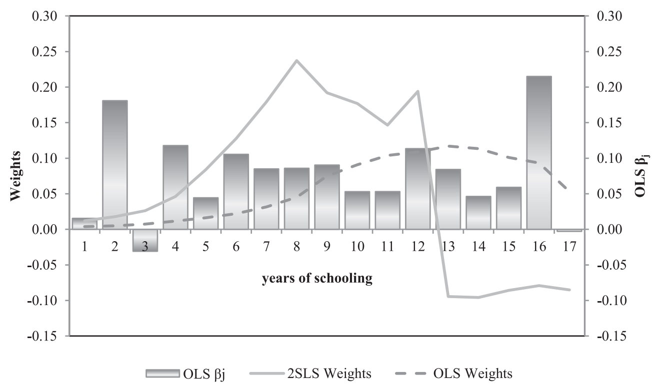


Finally, in the bottom panel, we turn to estimates of the private return to schooling using three dummies for compulsory schooling as instruments. While this analysis is based on that of Acemoglu and Angrist (2001), we consider the effects of schooling on log annual earnings rather than weekly wages for white men in their 40s. Figure 5 reports the OLS estimates of the β_j parameters as well as the OLS and 2SLS weights. OLS estimates indicate that an additional year of schooling translates into an 8.2 percent increase in annual earnings, while the 2SLS estimates suggest a much larger return. The reweighted OLS estimates fall in between the OLS and 2SLS estimates, although they are much closer to the OLS estimates. The effect of reweighting is minor despite substantially different OLS and 2SLS weights. Our test rejects the hypothesis that the reweighted OLS and 2SLS estimates are equal, even though the instruments are not particularly

[23] The standard error for this reweighted effect is derived using the delta method and the estimated covariance matrix $\hat{V}$ defined in section III.

[24] See online appendix table C1 for coefficient estimates and their standard errors. While it is possible that some $\hat{\beta}_j$ are biased upward and others downward so as to perfectly offset when the 2SLS weights are applied, this seems highly unlikely given the economics of the problem (see, e.g., Lochner & Moretti, 2004).

FIGURE 5.—EFFECTS OF SCHOOLING ON LOG ANNUAL EARNINGS FOR MEN (OLS ESTIMATES AND WEIGHTS)



strong in this application (the first-stage F-statistic for the instruments is only 29.5).

V. Conclusion

In applied work, OLS and IV estimates often differ. In many cases, the sign of the difference is surprising given economic theory and plausible assumptions about the direction of endogeneity bias. Influential work by Imbens & Angrist (1994), Angrist & Imbens (1995), and Heckman and Vytlacil (2005) has clarified the interpretation of IV estimates as a local average treatment effect when the regression parameter of interest varies across individuals. Our work complements the existing understanding of the differences between IV and OLS estimates when the model is misspecified.

We consider a specific class of models with a single finite-valued discrete endogenous regressor, exogenous regressors that are additively separable and enter linearly, and coefficients that do not vary across individuals. Models of this type are widely used in empirical research to study the effects of multivalued program treatments, drug dosage levels, and schooling attainment. We focus attention on the possibility that per unit treatment effects vary across levels of treatment.

The growing focus on identification of causal effects in economics has led many researchers to estimate models of this type using IV methods. Yet due to the limited availability of valid instruments, it is common to estimate models that assume uniform per unit treatment effects even when those effects are likely to vary across treatment levels, as frequently suggested by more general specifications estimated using OLS. We show that in this case, OLS and IV/2SLS estimators identify different weighted averages of all per unit effects, which can lead to incorrect conclusions about endogeneity when using a standard Hausman test.[25]

[25] Other important concerns include the strength and exogeneity of the instrument(s). Our approach abstracts from instrument-related problems' instead addressing problems associated with misspecification in the structural equation if at least one exogenous instrument is available.

The main contribution of this paper is to develop a simple generalization of the Hausman test to assess whether differential weighting and variable per unit treatment effects can explain the difference between OLS and IV/2SLS estimators. Within the class of models under consideration, this serves as a specification test for exogeneity under reasonable conditions. Conveniently, this test requires only a single instrument, making it useful in many applications.

REFERENCES

Acemoglu, Daron, and Joshua D. Angrist, "How Large Are Human-Capital Externalities? Evidence from Compulsory-Schooling Laws" (Vol. 15, pp. 9–74) *NBER Macroeconomics Annual 2000*, (Cambridge, MA: National Bureau of Economic Research, 2001).

Angrist, D. Joshua, Kathryn Graddy, and Guido W. Imbens, "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies* 67 (2000), 499–527.

Angrist, D. Joshua, and Guido W. Imbens, "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *JASA* 90 (1995), 431–442.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in Louis Christofides, E. Kenneth Grant, and Robert Swindinsky, eds., *Aspects of Labour Economics: Essays in Honour of John Vanderkamp* (Toronto: University of Toronto Press, 1995).

——— "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David Card, eds, *Handbook of Labor Economics*, vol. 3A (Amsterdam: Elsevier, 1999).

Carneiro, Pedro, James Heckman, and Edward Vytlacil, "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica* 78 (2010), 377–394.

Currie, Janet, and Enrico Moretti, "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *Quarterly Journal of Economics* 118 (2003), 1495–1532.

Guggenberger, Patrik, "The Impact of a Hausman Pretest on the Asymptotic Size of a Hypothesis Test," *Econometric Theory* 26 (2010), 369–382.

Hausman, J. A., "Specification Tests in Econometrics," *Econometrica* 46 (1978), 1251—1271.

Heckman, James J., and Edward Vytlacil, "Instrumental Variables Methods for the Correlated Random Coefficient Model," *Journal of Human Resources* 33 (1998), 974–987.

——— "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences* 96 (1999), 4730–4734.

——— "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73 (2005), 669–738.

Heckman, James J., Serglo Urzua, and Edward Vytlacil, "Understanding Instrumental Variables in Models with Essential Heterogeneity," this REVIEW 88 (2006), 389–432.

Imbens, Guido, and Joshua Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62 (1994), 467–475.

Kling, Jeffrey, "Interpreting Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business Economics and Statistics* 19 (2001), 358–364.

Lochner, Lance, "Nonproduction Benefits of Education: Crime, Health, and Good Citizenship," in E. Hanushek, S. Machin, and L. Woessmann, eds., *Handbook of the Economics of Education*, vol. 4 (Amsterdam: Elsevier Science, 2011).

Lochner, Lance, and Enrico Moretti, "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests and Self-Reports," NBER working paper 8605 (2001).

——— "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests and Self-Reports," *American Economic Review* 94 (2004), 155–189.

Moffitt, Robert, "Estimating Marginal Treatment Effects in Heterogeneous Populations," *Annales d'Economie et de Statistique* 91 (2009), 239–261.

Mogstad, Magne, and Matthew Wiswall, "Linearity in Instrumental Variables Estimation: Problems and Solutions," IZA discussion paper 5216 (2010).

White, Halbert, "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association* 76 (1981), 419–433.

Wong, Ka-fu, "Effects on Inference of Pretesting the Exogeneity of a Regressor," *Economic Letters* 56 (1997), 267–271.

Wooldridge, Jeffrey, "On Two Stage Least Squares Estimation of the Average Treatment Effect in Random Coefficient Models," *Economics Letters*, 56 (1997), 129–133.

Yitzhaki, Shlomo, "On Using Linear Regressions in Welfare Economics," *Journal of Business and Economic Statistics* 14 (1996), 478–486.