Identification and Inference with Many Invalid Instruments*

Michal Kolesár† Raj Chetty‡ John Friedman§ Edward Glaeser¶
Guido W. Imbens‖

October 2014

Abstract

We study estimation and inference in settings where the interest is in the effect of a potentially endogenous regressor on some outcome. To address the endogeneity we exploit the presence of additional variables. Like conventional instrumental variables, these variables are correlated with the endogenous regressor. However, unlike conventional instrumental variables, they also have direct effects on the outcome, and thus are "invalid" instruments. Our novel identifying assumption is that the direct effects of these invalid instruments are uncorrelated with the effects of the instruments on the endogenous regressor. We show that in this case the limited-information-maximum-likelihood (liml) estimator is no longer consistent, but that a modification of the bias-corrected two-stage-least-squares (tsls) estimator is consistent. We also show that conventional tests for over-identifying restrictions, adapted to the many instruments setting, can be used to test for the presence of these direct effects. We recommend that empirical researchers carry out such tests and compare estimates based on liml and the modified version of bias-corrected tsls. We illustrate in the context of two applications that such practice can be illuminating, and that our novel identifying assumption has substantive empirical content.

Keywords: Instrumental Variables, Misspecification, Many Instruments, Two-Stage-Least-Squares, Limited-Information-Maximum-Likelihood.

*Financial support for this research was generously provided through NSF grants 0820361 and 0961707. We are grateful for comments by participants in the econometrics lunch seminar at Harvard University, the Harvard-MIT Econometrics seminar, the NBER 2011 Summer Institute, the Oslo 2011 Econometric Society meeting, and in particular for discussions with Gary Chamberlain and Jim Stock, and comments by Caroline Hoxby and Tiemen Woutersen. We also thank the editor and two anonymous reviewers for valuable comments and suggestions.

†Department of Economics and Woodrow Wilson School, Princeton University. Electronic correspondence: mkolesar@princeton.edu.

‡Department of Economics, Harvard University, and NBER. Electronic correspondence: chetty@fas.harvard.edu.

§Kennedy School of Government, Harvard University, and NBER. Electronic correspondence: john_friedman@harvard.edu.

¶Department of Economics, Harvard University, and NBER. Electronic correspondence: eglaeser@harvard.edu.

‖Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu.

1 Introduction

In this paper we study estimation and inference in settings where the interest is in the effect of a potentially endogenous regressor on some outcome. To allow for the possible endogeneity, we exploit the presence of additional variables. These variables have some of the features of conventional instrumental variables, in the sense that they are correlated with the endogenous regressor. However, in contrast to conventional instrumental variables, these variables potentially also have direct effects on the outcome, and thus are "invalid" instruments.

Motivated by the context of our applications we explore the identifying power of a novel assumption that the direct effects of these invalid instruments are uncorrelated with the effects of the instruments on the endogenous regressor. We focus on the case with many instruments, allowing their number to increase in proportion with the sample size as in Kunitomo (1980), Morimune (1983) and Bekker (1994). To accommodate the structure in our applications in which the number of instruments is tied to the number of exogenous covariates, we also allow the number of exogenous covariates to increase in proportion with the sample size, as in Anatolyev (2011).

We show that the limited-information-maximum-likelihood (liml) estimator is no longer consistent once direct effects are present. On the other hand, the modified-bias-corrected-two-stage-least-squares (mbtsls) estimator remains consistent. This estimator is a modification of the bias-corrected two stage least squares estimator (Nagar, 1959; Donald and Newey, 2001) that allows for many exogenous covariates. The intuition for this result is that the liml estimator attempts to impose proportionality of *all* the reduced form coefficients. On the other hand mbtsls, like the two-stage least squares (tsls) estimator, can be thought of as a two-stage estimator. In the first stage a single instrument is constructed as a function of only instruments and endogenous regressors, not involving the outcome variable. This constructed instrument is then used in the second stage to estimate the parameter of interest using methods for just-identified settings. Identification only requires validity of the constructed instrument, not of all the individual instruments. The robustness of the mbtsls estimator comes at a price: the estimator is less efficient than liml in the absence of these direct effects under Normality and homoskedasticity.

We also show that conventional tests for over-identifying restrictions, adapted to the many

instruments setting, can be used to test for the presence of these direct effects. We recommend in practice that researchers carry out such tests and compare estimates based on liml and the modified version of bias-corrected tsls. We illustrate in the context of two applications that such practice can be illuminating.

The paper is related to two strands of literature. First, we contribute to the literature on many and weak instruments, started by Kunitomo (1980), Morimune (1983), Bekker (1994), Staiger and Stock (1997), and Chao and Swanson (2005). In recent work Anatolyev (2011) relaxes the assumption of fixed number of exogenous regressors. Hausman, Newey, Woutersen, Chao and Swanson (2012); Chao, Swanson, Hausman, Newey and Woutersen (2012) and Ackerberg and Devereux (2009) relax the assumption of homoscedasticity. Hansen, Hausman and Newey (2008), Belloni, Chen, Chernozhukov and Hansen (2012) and Gautier and Tsybakov (2011) allow the first stage to be estimated non-parametrically. This paper takes a complementary approach: we relax the assumption of no direct effects, but keep the rest of the model simple to maintain tractability. Our key contribution is to show that the superiority of liml in the homoscedastic Normal error case with many instruments is tied to the assumption of no direct effects. The mbtsls estimator is shown to be less efficient than liml in the case with no direct effects, but robust to the presence of uncorrelated direct effects.

Second, we contribute to the literature studying properties of instrumental variables methods allowing for direct effects of the instruments. This literature has largely focused on the case with a fixed number of instruments. The focus of this literature has been on correcting size distortions of tests, biases of estimators, sensitivity analyses, and bounds in the presence of direct effects. Fisher (1961, 1966, 1967), Caner (2007); Berkowitz, Caner and Fang (2008) and Guggenberger (2012) analyze the implications of local (small) violations of exogeneity assumption. Hahn and Hausman (2005) compare biases for different estimators in the presence of direct effects. Conley, Hansen and Rossi (2012); Ashley (2009) and Kraay (2008) propose sensitivity analyses in the presence of possibly invalid instruments. Nevo and Rosen (2012) consider assumptions about the sign of the direct effects of the instruments on the outcome to derive bounds on the parameters of interest. Reinhold and Woutersen (2011) and Flores and Flores-Lagunes (2010) also derive bounds allowing

for direct effects of the instruments on the outcome. The current paper is the first to derive (point) identification results in the presence of non-local departures from the no-direct-effects assumption or exclusion restriction.

The rest of the paper is organized as follows. In Section 2 we discuss in detail the empirical setting that motivates our study, based on Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011). In Section 3 we set up the general problem and formulate the critical assumptions. Next, in Section 4 we report on the large sample properties of k-class estimators, which covers both liml and mbtsls. In Section 5 we discuss tests for instrument validity. We then analyze two data sets to illustrate the usefulness of the results in Section 6. In Section 7 we report the results of a small simulation study to assess the accuracy of our asymptotic approximations. Section 8 concludes.

2 Motivating example

In this section we discuss the empirical application that motivates our setup. The application is based on Chetty *et al.* (2011). Chetty *et al.* (2011) are interested in estimating the effect of early achievement for children, as measured by kindergarten performance, on subsequent outcomes, say first grade scores. Chetty *et al.* (2011) wish to exploit the fact that kindergarten teachers are randomly assigned to classes, generating arguably exogenous variation in kindergarten performance. This suggests using kindergarten teacher or classroom indicators as instruments for kindergarten performance. However, a concern with this strategy is that classes mostly stay together over multiple years during the child's education. As a result, kindergarten classroom/teacher assignment is almost perfectly correlated with first grade classroom/teacher assignment. Therefore, the instrument (kindergarten teacher assignment) may have direct effects on the outcome (first grade performance) through first grade teacher assignment, that is, not mediated through the endogenous regressor (kindergarten performance). However, if first grade teachers are randomly assigned, and thus independent of kindergarten teacher assignment, the direct effect of the instrument on the outcome might reasonably be assumed to be independent of the direct effect of the instrument on the endogenous regressor. We show that this independence assumption has substantial identifying power, and discuss estimation strategies that exploit it. The identifying power of this independence

assumption suggests that in applications where there is concern regarding the presence of direct effects of the instruments on the outcome it may be useful to explore whether the substantive argument for their presence also suggests that these effects are independent of the effect of the instrument on the endogenous regressor.

To make this precise, let us discuss a simplified version of the Chetty *et al.* (2011) application in more detail. Let us ignore the presence of any exogenous regressors beyond the intercept. Children are indexed by $i = 1, \ldots, N$. The classroom or cluster variable is $G_i \in \{1, 2, \ldots, N_G\}$, where N_G is the number of clusters or classrooms. For simplicity, let us assume here that the classrooms are equal size. The instruments are the classroom indicators, $Z_{ik} = \mathbf{1}_{G_i=k}$, for $k = 1, \ldots, N_G - 1$, so that the number of instruments is the number of clusters minus one. Following the clustering literature we focus on large sample approximations where the number of units in each cluster is finite and the number of clusters increases proportional to the sample size, $N_G/N \to \alpha_K > 0$, leading to the Bekker-style many-instruments asymptotics. In this simple case the model can be written as

$$Y_i = \delta + \beta X_i + \sum_{k=1}^{N_G-1} \gamma_k Z_{ik} + \epsilon_i, \tag{2.1}$$

$$X_i = \pi_2 + \sum_{k=1}^{N_G-1} \pi_{1,k} Z_{ik} + \nu_i, \tag{2.2}$$

where Y_i is the outcome (first grade test scores) and X_i is the endogenous regressor (kindergarten performance). The residuals ϵ_i and ν_i are assumed to be independent across individuals, but correlated with each other. The coefficient β on the endogenous regressor is the object of interest. The coefficients on the instruments in the second equation, $\pi_{1,k}$ capture the direct effects on the endogenous regressor Here they represent the effects of the kindergarten teachers on kindergarten performance (we normalize $\pi_{1,N_G} = 0$, so that the effects are relative to classroom N_G). The presence of nonzero coefficients on the instruments in the first equation, denoted by γ_k, is what make the instruments invalid. These coefficients represent the effects of the first grade teachers on the first grade test scores.

Similar to the clustering literature, we can view the coefficients $\pi_{1,k}$ and γ_k as random variables.

In this setting where the instruments are cluster indicators this is equivalent to viewing the cluster effects as random, a common assumption in such settings. An alternative formulation of the model, one which stresses the links to the clustering literature, would be

$$Y_i = \delta + \beta X_i + U_{G_i} + \epsilon_i, \tag{2.3}$$

$$X_i = \pi_2 + V_{G_i} + \nu_i. \tag{2.4}$$

The random classroom component in the outcome equation in the clustering notation, U_{G_i}, is equal to the coefficient on one of the instruments, γ_{G_i}, and the random classroom component in the equation for X_i, V_{G_i} is equal to the coefficient on the same instrument in the first stage, π_{1,G_i}. The V_{G_i} represents the effect of kindergarten teachers on the kindergarten performance. The U_{G_i} represents the effect of first grade teachers on the outcome. We focus on the notation and formulation in (2.1)–(2.2) because it stresses links to the literature on many instrumental variables that are helpful in motivating the estimators we consider.

The instruments are not valid in the sense that the standard orthogonality condition for instruments does not hold, holding fixed the $\gamma_1, \ldots, \gamma_{N_G-1}$:

$$\begin{aligned}&\mathbb{E}\left[(Y_i - \delta - X_i\beta)Z_i | Z_i, \gamma_1, \ldots, \gamma_{N_G-1}\right] \\ &\qquad = \mathbb{E}\left[U_{G_i} Z_i |\, Z_i, \gamma_1, \ldots, \gamma_{N_G-1}\right] = \begin{pmatrix} \gamma_1 Z_{i1} \\ \vdots \\ \gamma_{N_G-1} Z_{iN_G-1} \end{pmatrix} \neq 0. \end{aligned} \tag{2.5}$$

However, we wish to exploit the random assignment of both kindergarten and first grade teachers. We therefore consider the assumption that the effects of kindergarten teachers on kindergarten performance and the effects of first grade teachers on outcomes are independent,

$$\pi_{1,k} \perp\!\!\!\perp \gamma_k,$$

or, given a normalization of the mean of the γ_k, $\mathbb{E}[\pi_{1,k}\gamma_k] = 0$. In terms of the cluster formulation (2.3), the assumption is $U_{G_i} \perp\!\!\!\perp V_{G_i}$. This suggests replacing the orthogonality condition (2.5),

which requires each instrument to be valid, with

$$\begin{aligned}\mathbb{E}\left[(Y_i-\delta-X_i\beta)Z_i'\pi_1\right] &= \mathbb{E}\left[\mathbb{E}\left[\left.(Y_i-\delta-X_i\beta)Z_i'\pi_1\right| Z_i\right]\right] \\ &= \mathbb{E}\left[U_{G_i}V_{G_i}\right] = \mathbb{E}\left[\sum_{k=1}^{N_G-1}\gamma_k\pi_{1,k}Z_{ik}\right] = 0,\end{aligned} \tag{2.6}$$

which requires the instruments to be valid in an average sense. Here π_1 is the vector with kth element equal to $\pi_{1,k}$. In a setting with a few instruments this would suggest estimating β as the solution to solving (2.6) with π_1 replaced by the least squares estimator $\hat{\pi}_1$:

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}-\beta(X_i-\overline{X})\right)Z_i'\hat{\pi}_1 = 0,$$

where $\overline{Y}$ and $\overline{X}$ are sample averages of Y_i and X_i respectively. Solving this for β leads to the standard tsls estimator. However, since the work by Bekker (1994) it is well known that even with valid instruments the tsls estimator is not consistent in settings with many instruments, and thus it is unlikely to be consistent here. This motivates looking for alternative, tsls-like, estimators of the type that have been proposed to deal with many-instrument problems. We do so in the Section 4. First, in Section 3, we introduce the general setup.

3 General setup

We consider the following instrumental variables model:

$$\begin{aligned}Y_i &= X_i\beta + W_i'\delta + Z_i'\gamma + \epsilon_i. \\ X_i &= Z_i'\pi_1 + W_i'\pi_2 + \nu_i.\end{aligned} \tag{3.1}$$

The first equation relates a scalar outcome Y_i, $i = 1,\ldots,N$, to a potentially endogenous scalar regressor X_i. W_i is a vector of exogenous regressors with dimension L_N (including an intercept), and Z_i is a vector of instruments with dimension K_N. The second equation relates the endogenous regressor X_i to the exogenous regressors W_i and the instruments Z_i. The object of interest is the coefficient β on the endogenous regressor in the outcome equation.

The model (3.1) modifies the conventional many-instruments model (e.g. Bekker, 1994) in two

ways. First, and this is the main contribution of the paper, we allow γ to be non-zero, thus allowing for direct effects of the instrument on the outcome. If we restrict $\gamma = 0$, then the exclusion restriction holds, and the instruments are valid. If we leave γ unrestricted, then β, the coefficient of interest, is not identified. In this paper, we will consider assumptions on γ that are weaker than $\gamma = 0$, but that still allow us to identify β, and assess their empirical content. Second, like Anatolyev (2011), we allow the number of exogenous regressors, L_N, to change with the sample size. The motivation for this extension is that often the presence of a large number of instruments is the result of interacting a few basic instruments with many exogenous covariates. For example, in Angrist and Krueger (1991), the basic instruments were three quarter of birth indicators. These were interacted with year of birth and state of birth indicators to generate a large number of instruments. As the results below show, this second extension does not make a substantial difference for the variance calculations, unless the ratio of the number of exogenous variables to the sample size is large. It does, however, matter for tests of instrument validity, as we discuss in Section 5.

Because the number of instruments and the number of exogenous variables change with the sample size, the distribution of some of the random variables also changes with the sample size. To be precise, we should therefore index the random variables and parameters by the sample size N. For ease of notation we drop this index. In the remainder γ and π_1 will be vectors of dimension K_N, and δ and π_2 will be vectors of dimension L_N.

Next, we introduce some additional notation. Let $\mathbf{Y}$ be the N-component vector with ith element Y_i, $\mathbf{X}$ the N-component vector with ith element X_i, ϵ the N-component vector with ith element ϵ_i, ν the N-component vector with ith element ν_i, $\mathbf{W}$ the $N \times L_N$ matrix with ith row equal to W_i', and $\mathbf{Z}$ the $N \times K_N$ matrix with ith row equal to Z_i'. Let $\overline{\mathbf{Z}} = (\mathbf{Z}, \mathbf{W})$ be the full matrix of exogenous variables. For an arbitrary $N \times J$ matrix $\mathbf{S}$, let

$$\mathbf{P_S} = \mathbf{S}\left(\mathbf{S'S}\right)^{-1}\mathbf{S'} \qquad \text{and} \qquad \mathbf{M_S} = \mathbf{I} - \mathbf{S}\left(\mathbf{S'S}\right)^{-1}\mathbf{S'}$$

denote the $N \times N$ projection matrix and the $N \times N$ annihilator matrix that projects on the orthogonal complement of $\mathbf{S}$.

Following Staiger and Stock (1997), we use the subscript $\perp$ as shorthand for taking residuals

after regression on the exogenous regressors $\mathbf{W}$, so $\mathbf{Z}_\perp = \mathbf{M}_\mathbf{W}\mathbf{Z}$, $\mathbf{X}_\perp = \mathbf{M}_\mathbf{W}\mathbf{X}$, and $\mathbf{Y}_\perp = \mathbf{M}_\mathbf{W}\mathbf{Y}$. We also denote by ι_N the N-dimensional vector of ones.

Define the augmented concentration parameter, the two by two matrix Λ_N:

$$\Lambda_N = \begin{pmatrix} \Lambda_{N.11} & \Lambda_{N,12} \\ \Lambda_{N,12} & \Lambda_{N,22} \end{pmatrix} = \begin{pmatrix} \gamma & \pi_1 \end{pmatrix}' \mathbf{Z}_\perp' \mathbf{Z}_\perp \begin{pmatrix} \gamma & \pi_1 \end{pmatrix}. \tag{3.2}$$

The (1,1) element, $\Lambda_{N,11}$, measures the degree of misspecification. In the case with valid instruments, $\gamma = 0$, and thus $\Lambda_{N,11} = \Lambda_{N,12} = 0$ and the only non-zero element of Λ_N is $\Lambda_{N,22}$. The $(2,2)$ element $\Lambda_{N,22}$ is closely related to the conventional concentration parameter (Mariano, 1973; Rothenberg, 1984), typically defined as $\Lambda_{N,22}/\Sigma_{22}$. Following Andrews, Moreira and Stock (2006), we use the version without dividing by the structural variance Σ_{22}.

We make the following assumptions. Some of these can be weakened along the lines of Chao and Swanson (2005). We focus on the simplest version of the assumptions and results that allow us to focus on the conceptual contribution of the paper.

Assumption 1 (Instruments and exogenous variables)**.**

(i) *$Z_i \in \mathbb{R}^{K_N}$, $W_i \in \mathbb{R}^{L_N}$, $\epsilon_i \in \mathbb{R}, \nu_i \in \mathbb{R}$, for $i = 1, \ldots, N$, $N = 1, \ldots$ are triangular arrays of random variables with $(Z_i, W_i, \epsilon_i, \nu_i)$, $i = 1, \ldots, N$ exchangeable.*

(ii) *$\overline{\mathbf{Z}}$ is full column rank with probability one.*

This assumption is standard, with a minor adaption to allow for many exogenous variables.

Assumption 2 (Model)**.**

(i) *$(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ are i.i.d. with mean zero, positive definite covariance matrix Σ, and finite fourth moments;*

(ii) *The distribution of $(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ is Normal.*

To simplify the derivation of distributional results, we will assume that the structural errors Normally distributed. We do not require Normality for consistency arguments. Recent papers by Chao *et al.* (2012) and Hausman *et al.* (2012) investigate the implications of heteroscedasticity in the setting with many valid instruments, and show that liml loses some of its attractive properties in

that case. Our results complement theirs in the sense that our results highlight a different concern with conventional estimators such as liml.

Assumption 3 (Number of instruments and exogenous regressors)**.** *For some* $0 \leq \alpha_K < 1$ *and* $0 \leq \alpha_L < 1$*, such that* $\alpha_K + \alpha_L < 1$

$$K_N/N = \alpha_K + o(N^{-1/2}), \qquad \text{and} \qquad L_N/N = \alpha_L + o(N^{-1/2}).$$

The first part of this assumption is standard in the many-instrument literature. The second part relaxes the standard assumption that the number of exogenous covariates is fixed and is identical to the corresponding assumption in Anatolyev (2011).

Assumption 4 (Concentration parameter)**.** *For some positive semi-definite* 2×2 *matrix* Λ *with* $\Lambda_{22} > 0$*,*

$$\Lambda_N/N \xrightarrow{p} \Lambda, \qquad \text{and} \qquad \mathbb{E}\left[\Lambda_N/N\right] \to \Lambda.$$

The first part of Assumption 4 is a natural extension of the assumption underlying the Bekker many-instrument asymptotics. The second part of the assumption strengthens this slightly by also requiring the expectation of the augmented concentration parameter to converge to its probability limit.

Assumption 5 (Zero Correlation)**.** $\Lambda_{12} = 0$*.*

The last assumption is a new and critical assumption. We allow for direct effects of the instruments on the outcome ($\Lambda_{11} > 0$), but assume that these direct effects are orthogonal to the direct effects of the instruments on the endogenous regressor. This is a strong assumption, and one that needs to be justified on a case-by-case basis. In settings such as the Chetty *et al.* (2011) application we argued (in Section 2) that this may be a reasonable assumption.

Typically, this assumption will implicitly require that the number of instruments increase with the sample size. In the Chetty *et al.* (2011) application, for instance, $\Lambda_{12,N}/N$ equals the sample

correlation between the kindergarten teacher effect $\pi_{1,k}$ and the first-grade teacher effect γ_k,

$$\frac{1}{N_G}\sum_{k=1}^{N_G-1}\pi_{1,k}\gamma_k - \frac{1}{N_G}\sum_{k=1}^{N_G-1}\pi_{1,k}\cdot\frac{1}{N_G}\sum_{k=1}^{N_G-1}\gamma_k.$$

With a fixed number of clusters, $\Lambda_{12,N}/N$ will in general not average out to exactly zero unless $\gamma = 0$. However, if first-grade and kindergarten teachers are assigned independently, so that the population correlation between $\pi_{1,k}$ and γ_k is zero, $\Lambda_{12,N}/N$ will converge to zero as the number of clusters $N_G \to \infty$. The requirement that the number of instruments/clusters increases with the sample size is similar to that in the clustering literature, in which the number of clusters needs to increase with the sample size to achieve point-identification.

In some applications, there may therefore potentially be a bias-variance trade-off as adding instruments both decreases the bias coming from $\Lambda_{12,N}/N$ not being equal to zero in the sample, and decreases the precision of the estimates due to the many instrument problem. For instance, when the instruments are group indicators (as in the Chetty *et al.* (2011) application), for a fixed total sample size, there is a trade-off between getting data on more clusters with less observations per cluster, and getting data on less clusters, but more observations per cluster.

4 The Properties of k-Class Estimators

This section contains the main formal results of the paper. We discuss estimators for β and their large sample properties under the assumptions introduced in the previous section. Some of the results are for general k-class estimators (Nagar, 1959; Theil, 1961, 1971; Davidson and MacKinnon, 1993), and some for four particular estimators in this class. All four have been introduced previously, and are asymptotically equivalent in the conventional setting with a fixed number of valid instruments and a fixed number of exogenous regressors. Given a scalar κ, a k-class estimator for (β, δ) is given by:

$$\begin{pmatrix}\hat{\beta}_\kappa\\ \hat{\delta}_\kappa\end{pmatrix} = \left(\begin{pmatrix}\mathbf{X} & \mathbf{W}\end{pmatrix}'(\mathbf{I}-\kappa\cdot\mathbf{M}_{\overline{\mathbf{Z}}})\begin{pmatrix}\mathbf{X} & \mathbf{W}\end{pmatrix}\right)^{-1}\left(\begin{pmatrix}\mathbf{X} & \mathbf{W}\end{pmatrix}'(\mathbf{I}-\kappa\cdot\mathbf{M}_{\overline{\mathbf{Z}}})\mathbf{Y}\right).$$

We are primarily interested in the estimator for β, which can be written using the $\perp$ projection notation as

$$\hat{\beta}_\kappa = \left(\mathbf{X}_\perp'(\mathbf{I} - \kappa \cdot \mathbf{M}_{\mathbf{Z}_\perp})\mathbf{X}_\perp\right)^{-1}\left(\mathbf{X}_\perp'(\mathbf{I} - \kappa \cdot \mathbf{M}_{\mathbf{Z}_\perp})\mathbf{Y}_\perp\right). \tag{4.1}$$

A prominent member of the k-class is the two-stage-least-squares (tsls) estimator (Basmann, 1957; Theil, 1961), with $\hat{\kappa}_{\text{tsls}} = 1$. Even if all instruments are valid, this estimator has been shown to be inconsistent under many-instrument asymptotics, see Kunitomo (1980) and Bekker (1994). We also consider a bias-corrected version of the tsls estimator that is valid under many-instrument asymptotics. Nagar (1959) suggested the bias correction $\hat{\kappa}_{\text{nagar}} = 1 + (K_N - 2)/N$, but the second of the four estimators we focus on is a slightly different version suggested by Donald and Newey (2001), with

$$\hat{\kappa}_{\text{btsls}} = \frac{1}{1 - (K_N - 2)/N}.$$

Although in samples with a moderate number of instruments the difference between the Nagar and Donald-Newey estimators is small, this difference does not go away under many-instruments asymptotics with $K_N/N \to \alpha_K > 0$, and only the Donald-Newey version is consistent under those asymptotics. Once we also allow L_N to increase with sample size, $\hat{\beta}_{\text{btsls}}$ is no longer consistent. To address this issue, the third estimator we consider is a further modification of the Donald-Newey bias-corrected estimator, first suggested by Anatolyev (2011), that is consistent even when $L_N/N \to \alpha_L > 0$:

$$\hat{\kappa}_{\text{mbtsls}} = \frac{1 - L_N/N}{1 - K_N/N - L_N/N}.$$

In practice this modification has only a minor effect, unless the ratio of the number of exogenous variables to the sample size is substantial.

The fourth estimator we consider is the limited-information-maximum-likelihood (liml) estimator of Anderson and Rubin (1949), with

$$\hat{\kappa}_{\text{liml}} = \min_\beta \frac{(\mathbf{Y} - \mathbf{X}\beta)'\,\mathbf{M}_{\mathbf{W}}\,(\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)'\,\mathbf{M}_{\overline{\mathbf{Z}}}\,(\mathbf{Y} - \mathbf{X}\beta)}.$$

This estimator has been shown to be asymptotically efficient under many-instrument asymptotics (Chioda and Jansson, 2009; Anderson, Kunitomo and Matsushita, 2010) in the class of invariant estimators given Normality and homoskedasticity of the error terms.

The first of our two main results describes the probability limit of a general k-class estimator under the assumptions given in the previous section.

Theorem 1 (probability limits of k-class estimators)**.** *Suppose Assumptions 1, 2(i), 3, 4 and 5 hold. If* $\hat{\kappa} \xrightarrow{p} \kappa$ *with* $\kappa < \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\Lambda_{22}}{\Sigma_{22}(1-\alpha_K-\alpha_L)}$*, then:*

$$\hat{\beta}_{\hat{\kappa}} \xrightarrow{p} \beta_\kappa = \beta + \frac{(1-\alpha_L-(1-\alpha_K-\alpha_L)\kappa)\Sigma_{12}}{\Lambda_{22}+(1-\alpha_L-(1-\alpha_K-\alpha_L)\kappa)\Sigma_{22}}.$$

If we impose $\alpha_L = 0$, the condition for consistency of $\hat{\beta}_{\hat{\kappa}}$ is the same as in Chao and Swanson (2005), namely that $\hat{\kappa} \to 1/(1-\alpha_K)$. Having many exogenous regressors changes the condition on $\hat{\kappa}$ to $\hat{\kappa} \to (1-\alpha_L)/(1-\alpha_K-\alpha_L)$. As long as $\Lambda_{12} = 0$, this result holds whether or not $\Lambda_{11} > 0$. Therefore, the robustness of k-class estimators to the presence of direct effects depends on whether the probability limit of $\hat{\kappa}$ remains unaffected by their presence.

For the four estimators we discussed, the implication of this theorem is given in the following Corollary.

Corollary 1. *Suppose Assumptions 1, 2(i), 3, 4 and 5 hold. Then:*

(i) *(tsls)*

$$\beta_{tsls} = \beta + \frac{\alpha_K \Sigma_{12}}{\Lambda_{22}+\alpha_K\Sigma_{22}}, \qquad \kappa_{tsls} = 1,$$

(ii) *(btsls)*

$$\beta_{btsls} = \beta + \frac{\{\alpha_K\alpha_L/(1-\alpha_K)\}\,\Sigma_{12}}{\Lambda_{22}+\{\alpha_K\alpha_L/(1-\alpha_K)\}\,\Sigma_{22}}, \qquad \kappa_{btsls} = \frac{1}{1-\alpha_K},$$

(iii) *(mbtsls)*

$$\beta_{mbtsls} = \beta, \qquad \kappa_{mbtsls} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L},$$

(iv) *(liml) Suppose* $\min \text{eig}(\Sigma^{-1}\Lambda) < \Lambda_{22}/\Sigma_{22}$. *Then:*

$$\beta_{liml} = \beta - \frac{\min \text{eig}(\Sigma^{-1}\Lambda)\Sigma_{12}}{\Lambda_{22} - \min \text{eig}(\Sigma^{-1}\Lambda)\Sigma_{22}}, \qquad \kappa_{liml} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\min \text{eig}(\Sigma^{-1}\Lambda)}{1-\alpha_K-\alpha_L},$$

The key insight is that the mbtsls modification of the tsls estimator that makes it robust to the presence of many instruments and many exogenous variables is also robust to the presence of direct effects, provided these direct effects are uncorrelated with the effects of the instrument on the endogenous regressor. On the other hand, in order for liml to be consistent for all values of Σ, then it has to be the case that Λ_{11} is equal to zero since $\min \text{eig}(\Sigma^{-1}\Lambda) > 0$ otherwise. To provide some intuition, consider the reduced-form based on the model (3.1):

$$Y_i = Z_i'(\pi_1\beta + \gamma) + W_i'(\delta + \pi_2\beta) + (\nu_i\beta + \epsilon_i),$$
$$X_i = Z_i'\pi_1 + W_i'\pi_2 + \nu_i.$$

If the instruments are valid, so that $\gamma = 0$, then the vector of reduced-form coefficients on Z_i in the first equation is proportional to π_1, the vector of reduced-form coefficients in the second equation. The liml estimator tries to impose this proportionality. This leads to efficiency if proportionality holds, under Normality and homoskedasticity, (Chioda and Jansson, 2009; Anderson *et al.*, 2010). However, if $\gamma \neq 0$, then the proportionality does not hold in the population, and liml loses consistency. On the other hand, mbtsls, like tsls, can be thought of as two stage estimators. In the first stage composite instruments are constructed, one for each regressor (endogenous or exogenous) based on the data on the endogenous regressor, the exogenous variables, and the instruments alone. These instruments are then used to estimate the parameters of interest using a method for just-identified settings, possibly with some adjustment. In this procedure proportionality of the reduced forms is never exploited. This explains why $\Lambda_{12} = 0$ is a sufficient condition for consistency, although it results in efficiency loss relative to liml when proportionality does hold.

Note also that the bias of the btsls estimator is relatively minor: it is essentially proportional to the product of α_K and α_L, so that unless both are substantial, the bias will generally be small. However, the presence of many exogenous regressors might have a large effect on the probabil-

ity limits of other estimators. For example, in previous version of this paper (Kolesár, Chetty, Friedman, Glaeser and Imbens, 2011) we show that the jackknife instrumental variables estimator (Angrist, Imbens and Krueger, 1999) may exhibit substantial bias when the number of exogenous covariates is large.

Without the assumption that the direct effects are uncorrelated (Assumption 5), if follows from the proof of Theorem 1 that the probability limit of k-class estimators has an additional term that is proportional to Λ_{12}:

$$\hat{\beta}_{\hat{\kappa}} \xrightarrow{p} \beta + \frac{\Lambda_{12} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L)\kappa)\Sigma_{12}}{\Lambda_{22} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L)\kappa)\Sigma_{22}} \tag{4.2}$$

In this case all the k-class estimators are in general inconsistent, and in fact there are no estimators for β that are consistent for all values of Σ. On the other hand, the asymptotic bias of mbtsls, $\Lambda_{12}/\Lambda_{22}$, will be small so long as the covariance between the effect of the instruments on the outcome and their effect on the endogenous regressor is small relative to the strength of the instruments, Λ_{22}.

The second main result concerns the asymptotic approximation to the distribution of the mbtsls estimator. We focus on the mbtsls estimator because that is the only estimator in the k-class that is consistent under the assumptions we consider. A complication arises because, except in the special case where the only non-zero element of Λ_N is the $(2,2)$ element $\Lambda_{N,22}$ (the standard case with valid instruments, $\Lambda_{11} = 0$), the asymptotic distribution for $\hat{\beta}_{\text{mbtsls}}$ depends on the stochastic properties of $\Lambda_N - \Lambda$. In order to derive the asymptotic distribution of $\hat{\beta}_{\text{mbtsls}}$ we therefore make one additional assumption about the sequence of γ_k and $\pi_{1,k}$. That is, similar to corresponding assumptions in the clustering literature, we assume that these parameters are random and make assumptions regarding their stochastic properties. First we redefine the parameters by orthogonalizing them with respect to $\mathbf{Z}_\perp$ as

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_1 \end{pmatrix} = \left(\alpha_K \mathbf{Z}_\perp' \mathbf{Z}_\perp\right)^{1/2} \begin{pmatrix} \gamma & \pi_1 \end{pmatrix}$$

Assumption 6 (Incidental parameters)**.** *The pairs $(\tilde{\gamma}_k, \tilde{\pi}_{1,k})$, for $k = 1, 2, \ldots, K_N$, are i.i.d. with*

distribution

$$\left.\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{1,k} \end{pmatrix}\right| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}, \Xi\right).$$

The motivation for formulating the random effects assumption in terms of the orthogonalized parameters, rather than in terms of the original parameters, comes from the cluster structure in our example in Section 2. Exploiting that special structure, the augmented concentration parameter can be written as the sample covariance matrix of $(\gamma_k, \pi_{1,k})$:

$$\Lambda_N = \frac{N}{N_G} \sum_{k=1}^{N_G-1} \begin{pmatrix} (\gamma_k - \overline{\gamma})^2 & (\gamma_k - \overline{\gamma})(\pi_{1,k} - \overline{\pi}_{12}) \\ (\gamma_k - \overline{\gamma})(\pi_{1,k} - \overline{\pi}_{12}) & (\pi_{1,k} - \overline{\pi}_{12})^2 \end{pmatrix},$$

where

$$\overline{\gamma} = \frac{1}{N_G} \sum_{k=1}^{N_G-1} \gamma_k, \qquad \text{and} \qquad \overline{\pi}_{12} = \frac{1}{N_G} \sum_{k=1}^{N_G-1} \pi_{1,k}.$$

Now let us consider Assumption 6 and interpret it in this context. Suppose we have a large population of clusters. Let $\delta + U_k$ and $\pi_2 + V_k$ be the population means of $Y_i - \beta X_i$ and X_i in cluster k, and let δ and π_2 be the population average of the cluster means. In terms of the original parametrization, we have: $\pi_{1,k} = V_k$, and $\gamma_k = U_k$.

The natural way to impose a random effects structure on the parameters would be to assume that the cluster means $(\delta + U_k, \pi_2 + V_k)$ are independent and Normally distributed:

$$\begin{pmatrix} \delta + U_k \\ \pi_2 + V_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta \\ \pi_2 \end{pmatrix}, \Phi\right). \tag{4.3}$$

This implies

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_{1,k} \end{pmatrix} = \sqrt{\frac{N_G - 1}{N_G}} B \begin{pmatrix} \delta + U_1 & \pi_2 + V_1 \\ \vdots & \vdots \\ \delta + U_{N_G} & \pi_2 + V_{N_G} \end{pmatrix},$$

$$B = \left(\begin{array}{c|c} I_{N_G-1} - \frac{1-1/\sqrt{N_G}}{N_G-1} \iota_{N_G-1} \iota'_{N_G-1} & -\frac{1}{\sqrt{N_G}} \iota_{N_G-1} \end{array} \right)$$

where the $(N_G - 1) \times N_G$ matrix B satisfies $B\iota_{N_G} = 0$, and $BB' = \mathbf{I}_{N_G-1}$. Thus, a random effects specification on $(\delta + U_k, \pi_2 + V_k)$ as in (4.3) implies a random effects specification on $(\tilde{\gamma}, \tilde{\pi}_1)$, namely

$$\left.\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{1,k} \end{pmatrix}\right| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Xi\right), \qquad \text{with} \quad \Xi = \frac{N_G - 1}{N_G} \cdot \Phi. \tag{4.4}$$

Given Assumption 6, it follows that the augmented concentration parameter satisfies

$$\begin{aligned} \Lambda = \operatorname{plim}\left(\frac{\Lambda_N}{N}\right) &= \operatorname{plim}\left(\frac{1}{N}\begin{pmatrix} \gamma' \\ \pi_1' \end{pmatrix} \mathbf{Z}_\perp' \mathbf{Z}_\perp \begin{pmatrix} \gamma & \pi_1 \end{pmatrix}\right) \\ &= \operatorname{plim}\left(\frac{1}{K_N}\sum_{k=1}^{K_N}\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{1,k} \end{pmatrix}\begin{pmatrix} \tilde{\gamma}_k & \tilde{\pi}_{1,k} \end{pmatrix}\right) = \begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}\begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}' + \Xi. \end{aligned}$$

Now we can state the second main result of the paper.

Theorem 2 (Asymptotic Normality with many invalid instruments)**.** *Suppose that Assumptions 1–6 hold. Suppose, in addition, that $\alpha_K > 0$ and that $\Xi_{12} = 0$. Then*

$$\begin{aligned} &\sqrt{N}\left(\hat{\beta}_{mbtsls} - \beta\right) \xrightarrow{d} \\ &\qquad \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right) + \Lambda_{11}\left(\Sigma_{22} + \frac{\Lambda_{22}}{\alpha_K}\right)\right)\right). \end{aligned} \tag{4.5}$$

If in addition $\Lambda_{11} = 0$ (corresponding to the conventional many-instrument case), the distribution for $\hat{\beta}_{\text{mbtsls}}$ is the special case of (4.5) with $\Lambda_{11} = 0$:

$$\sqrt{N}\left(\hat{\beta}_{\text{mbtsls}} - \beta\right) \mid \overline{\mathbf{Z}} \xrightarrow{d} \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right). \tag{4.6}$$

In this case imposing Assumption 6 has no effect on the asymptotic distribution. This result obtains because under the standard valid many-instrument asymptotic sequence, the Normal prior on the incidental parameters gets dominated, and the Bernstein-von Mises theorem applies (see Kolesár, 2012).

The theorem assumes that the random effects are uncorrelated, $\Xi_{12} = 0$ to rule out the case in which $\tilde{\gamma}$ and $\tilde{\pi}_1$ are correlated, but their second moment is zero, $\Lambda_{12} = \Xi_{12} + \mu_\gamma\mu_\pi = 0$ because the

correlation happens to be exactly offset by the means. This assumption is moot when instruments are cluster indicators, because then the means μ_γ and μ_π can be normalized to zero, with the mean effect of the instruments being captured by the terms in π_2 and δ that correspond to the intercept, as in Equation (4.4).

The theorem also assumes Normality (Assumption 2(ii)). It is be possible to relax this assumption and instead only assume finite fourth moments (Assumption 2 (i)). Then the expression for asymptotic variance would have additional terms (Hansen *et al.*, 2008; van Hasselt, 2010). However, these additional terms will be small unless the distribution of the error terms displays substantial skewness or kurtosis *and* the design matrix of the instruments is unbalanced, so that there are observations with high leverage, $(\mathbf{P}_{\mathbf{Z}_\perp})_{ii}$ (Kolesár, 2012). We focus on the Normal case here to better highlight the substantive effect of relaxing the standard assumption that $\gamma = 0$.

The asymptotic variance of $\hat{\beta}_{\text{mbtsls}}$ is strictly larger if $\Lambda_{11} > 0$ than if $\Lambda_{11} = 0$. The additional term in the variance, $\Lambda_{11}\left(\Sigma_{22} + \Lambda_{22}/\alpha_K\right)$, diverges if α_K goes to zero. In contrast with much of the many-instruments literature, where the presence of many instruments is a nuisance, the number of instruments needs to increase with the sample size ($\alpha_K > 0$) for convergence of the estimator to be at $\sqrt{N}$ rate. The large number instruments is required so that $\Lambda_{12,N}/N$ converges to zero at $\sqrt{N}$ rate. This is similar to the clustering literature, in which the number of clusters needs to increase in proportion with the sample size.

One may be tempted to avoid this problem by not scaling by $\sqrt{\alpha_K}$ in the definition of $(\tilde{\gamma}, \tilde{\pi}_1)$. However, such scaling would be rather unusual. For instance, in the clustering example above, it follows from Equations (4.3) and (4.4) that for Assumption 6 to hold, the variance of the cluster effects U_k and V_k would need to be proportional to K_N/N: if $K_N/N \to 0$, this implies that in the limit, the cluster effects are exactly the same in each cluster, contradicting the Assumptions about Λ_N/N (Assumptions 4 and 5).

For comparison, the asymptotic distribution of liml given $\Lambda_{11} = 0$ is

$$\sqrt{N}\left(\hat{\beta}_{\text{liml}} - \beta\right) \mid \overline{\mathbf{Z}} \xrightarrow{d} \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right)\right)\right), \tag{4.7}$$

with a smaller variance than the mbtsls estimator under the same assumptions (comparing (4.6)

with (4.7)), consistent with the efficiency of liml under those conditions. There is therefore a trade-off between the robustness of the mbtsls estimator to the presence of direct (uncorrelated) effects and the efficiency of liml in the absence of such effects (under Normality and homoskedasticity).

5 Testing

The assumption that the instruments are valid (that is, that $\gamma = 0$) is equivalent to restricting the $\Lambda_{11,N}$ (and thus $\Lambda_{12,N}$) elements of the augmented concentration matrix to zero. Several tests of this restriction have been proposed in the literature, most of them in the setting with a fixed number of instruments, but some designed to be robust to the presence of many instruments.

The most popular one test, due to Sargan (1958), is based on the statistic:

$$J_{\text{Sargan}} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{liml}})'\mathbf{P}_{\mathbf{Z}_\perp}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{liml}})}{(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{liml}})'\mathbf{M}_{\mathbf{W}}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{liml}})/N} = N(1 - \hat{\kappa}_{\text{liml}}^{-1})$$

This statistic can easily be computed as the $N \cdot R^2$ from regressing the estimated residuals in the structural equation on instruments and exogenous regressors. Sargan (1958) shows that under the standard strong instrument asymptotic sequence which keeps the number of instruments and exogenous regressors fixed (so that $K_N = K$ and $L_N = L$), this statistic satisfies $J_{\text{Sargan}} \xrightarrow{d} \chi^2_{K-1}$. Anatolyev and Gospodinov (2011) show, however, that if the number of instruments is allowed to grow with the sample size, the limiting distribution is Normal, and using a critical value based on the χ^2 distribution with $K_N - 1$ degrees of freedom yields an asymptotically conservative test. Anatolyev and Gospodinov (2011) therefore propose an adjustment to the critical value. Unfortunately, if the number of exogenous regressors is allowed to grow with the sample size as well, the original as well as the adjusted Sargan test have asymptotic size equal to one (Anatolyev, 2011). We therefore propose to use a test statistic suggested by Cragg and Donald (1993):

$$J_{\text{Cragg-Donald}} = (N - K_N - L_N)(\hat{\kappa}_{\text{liml}} - 1)$$

Like the Sargan statistic, this statistic depends on the data only through $\hat{\kappa}_{\text{liml}}$. Both tests reject for large values of $\hat{\kappa}_{\text{liml}}$, so their power properties are identical; the only difference between them is in how well they control size. Under the standard strong instrument asymptotics, this statistic,

like the Sargan statistic, is also distributed according to χ^2_{K-1}. However, under many-instrument asymptotics, using the $1-\tilde{\alpha}$ quantile of the χ^2 distribution with (K_N-1) degrees of freedom for a test with nominal size $\tilde{\alpha}$ results in asymptotic size distortions. We therefore compare $J_{\text{Cragg-Donald}}$ against the $\Phi(\sqrt{(1-\alpha_L)/(1-\alpha_K-\alpha_L)}\Phi^{-1}(1-\tilde{\alpha}))$ quantile of $\chi^2_{K_N-1}$, where Φ is the cdf of a standard Normal distribution. Kolesár (2012) shows that this adjusted Cragg-Donald test controls size under strong, as well as many-instrument asymptotics.

6 Two Applications

In this section we discuss two applications. These will serve to provide further context for the empirical content of the assumptions, and in particular the zero correlation assumption (Assumption 5).

6.1 Application I

The first application is based on Chetty *et al.* (2011) first introduced in Section 2. The interest in Chetty *et al.* (2011) is in the effect of kindergarten performance on later outcomes. Here we focus on first, second, and third grade performance as the outcome of interest. The outcome equation is

$$Y_i = \beta X_i + \sum_{\ell=1}^{L_N} \delta_\ell W_{i\ell} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i. \tag{6.1}$$

Here the outcome Y_i is first, second, or third grade performance. The endogenous regressor X_i is kindergarten performance. The exogenous regressors W_{ik} include 76 school indicators and three demographic variables (female, black, and being on subsidized lunches), for a total of $L_N = 79$ exogenous variables. The instruments are $K_N = 238$ classroom indicators. The first stage is

$$X_i = \sum_{k=1}^{K_N} \pi_{1,k} Z_{ik} + \sum_{\ell=1}^{L_N} \pi_{2,\ell} W_{i\ell} + \nu_i. \tag{6.2}$$

The motivation for the zero correlation assumption is that the γ_k represent the effects of the first or subsequent, grade teachers. Because the classes largely stay the same from year to year, children with the same kindergarten teacher would have the same first, second, and third grade

teacher. However, by design the subsequent teachers were assigned randomly, independently of the kindergarten teachers, and so the γ_k would be independent of the $\pi_{1,k}$ if the only direct effect of the kindergarten classroom/teacher assignment was through the subsequent teacher.

Finally, we impose a random effects structure on the effects of the instruments on outcomes and endogenous regressors:

$$\left(\begin{array}{c}\tilde{\gamma}_k \\ \tilde{\pi}_{1,k}\end{array}\right)\Bigg|\,\mathbf{Z},\mathbf{W} \sim \mathcal{N}\left(\left(\begin{array}{c}0\\0\end{array}\right),\Lambda\right).$$

where, as before, the $(\tilde{\gamma}_k, \tilde{\pi}_{1,k})$ are the orthogonalized coefficients on the instruments:

$$\left(\begin{array}{cc}\tilde{\gamma} & \tilde{\pi}_1\end{array}\right) = \left(\alpha_K \mathbf{Z}_\perp'\mathbf{Z}_\perp\right)^{1/2}\left(\begin{array}{cc}\gamma & \pi_1\end{array}\right).$$

In Table 1 we present point estimates of the parameter of interest, β, based on tsls, liml, btsls, and mbtsls. For each of the estimators we present up to four different standard errors: conventional standard errors, Bekker standard errors which are robust to the presence of many instruments, standard errors robust to the presence of many instruments and many exogenous regressors, and standard errors robust to the presence of direct effects of the instruments on the outcome. For all three outcomes the liml estimate differ substantially from tsls. Based on the early many-instrument literature one might interpret that as evidence of the bias of the tsls estimator in settings with many instruments, and view the liml estimates are more credible. However, the btsls and mbtsls estimates, which, like liml, would be consistent under the conventional many-instruments asymptotics, also differ substantially from the liml estimates.

To understand the difference between the liml and btsls/mbtsls estimates, we report in Table 2 test statistics and p-values for the tests for instrument validity $\Lambda_{11} = 0$. The results from these tests are consistent with substantial variation in the γ_k. Although these results do not validate the mbtsls estimates (for that one still relies on the zero correlation assumption, $\Lambda_{12} = 0$), at the very least they imply that the liml estimates should not be taken at face value.

6.2 Application II

In the second application we apply some of the methods to a subset of the Angrist and Krueger (1991) data, who study the effect of years of schooling on log-earnings. We use interactions between quarter, year, and state of birth as instruments, and restrict the sample to individuals born in the first and fourth quarter (so we have a single binary basic instrument, although this is not essential), dropping observations from Alaska because there are some years birth quarters with no observations, leaving us with observations on 162,487 individuals.

Let W_{ik}, for $k = 1, \ldots, K_N$ be the cluster indicators, corresponding to year-of-birth times state-of-birth interactions, so that $K_N = 500$, and let Q_i be the binary quarter-of-birth indicator. The general model we consider is

$$Y_i = \beta X_i + \sum_{k=1}^{K_N} \delta_k W_{ik} + \sum_{k=1}^{K_N} \gamma_k Q_i W_{ik} + \epsilon_i, \tag{6.3}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{1,k} Q_i W_{ik} + \sum_{k=1}^{K_N} \pi_{2,k} W_{ik} + \nu_i, \tag{6.4}$$

where Y_i is log-earnings, and X_i is years of schooling. We assume a random effects structure

$$\left(\begin{array}{c} \gamma_k \\ \pi_{1,k} \end{array} \right) \Bigg| \mathbf{Z}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right), \Xi \right).$$

The critical assumption that $\Lambda_{12} = 0$ is more difficult to justify in this case than in the Chetty *et al.* (2011) case. Its plausibility relies on the interpretation of the direct effects of the instruments on the outcome and the endogenous regressor. The argument for the direct effects of the instrument on the endogenous regressor in the AK study is that quarter of birth effects years of schooling through compulsory schooling laws. If the direct effects of quarter of birth on earnings is through other differences between states, either in institutions or in economic climate, it may be reasonable to assume that these other differences are uncorrelated with compulsory schooling laws. However, unlike in the Chetty *et al.* (2011) study, there is no design feature that makes this assumption more plausible. Nevertheless, in our view it is still useful to calculate both liml and mbtsls, and calculating the p-value for the test of instrument validity. Finding that the estimators are similar,

and that the p-values are not unusually small, lends support to the instrumental variables estimates.

We report in Table 3 estimates for β based on tsls, liml, btsls, and mbtsls and the various standard errors. In Table 4 we report the results based on the Sargan and Craig-Donald tests for validity of instruments. Here we find, in contrast to the findings for the Chetty *et al.* (2011) data, that the three estimators, liml, btsls, and mbtsls are very similar, and that there is no evidence of direct effects of the instruments on the outcome. Note also that although L_N and K_N are equal in magnitude, the additional adjustment in moving from btsls to mbtsls again makes little difference.

7 A Simulation Study

We also carried out a small simulation study to assess the finite sample properties of the estimators. The design was based on the Chetty *et al.* (2011) study. The model is

$$Y_i = \beta X_i + \sum_{\ell=1}^{L_N} \delta_\ell W_{i\ell} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i, \tag{7.1}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{1,k} Z_{ik} + \sum_{\ell=1}^{L_N} \pi_{2,\ell} W_{i\ell} + \nu_i, \tag{7.2}$$

where $W_{i\ell}$ and Z_{ik} are school and classroom indicators from the Chetty *et al.* (2011) data, so that $L_N = 76$ and $K_N = 238$. We put a random effects structure on the orthogonalized parameters:

$$\left(\begin{array}{c} \tilde{\gamma}_k \\ \tilde{\pi}_{1,k} \end{array} \right) \Bigg| \, \mathbf{Z}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \Lambda \right).$$

The sample size in the simulations is $N = 4,170$, corresponding to the sample size in the Chetty *et al.* (2011) data, so that $\alpha_L = 0.0182$ and $\alpha_K = 0.0571$.

The values of the parameters are $\beta = 0$, $\delta_j = 0$ and $\pi_{2,j} = 0$, for $j = 1, \ldots, L_N$. The covariance matrix for the structural errors is

$$\Sigma = \left(\begin{array}{cc} 1.0 & 0.5 \\ 0.5 & 1.0 \end{array} \right).$$

The coefficients γ_k and $\pi_{1,k}$ are drawn from Normal distributions centered at zero and variances so

that

$$\Lambda_{11,N}/K_N = 0.7, \qquad \Lambda_{22,N}/K_N = 2.4,$$

comparable to the values from Chetty *et al.* (2011). We also consider $\Lambda_{11,N} = 0$.

For each of the four estimators we calculate the bias as the average difference between the estimate and the true value (note that liml does not have finite moments, so the bias is arguably not a useful summary measure), the median absolute deviation, and coverage rates based on confidence intervals using the four different standard errors: conventional standard errors, Bekker standard errors which are robust to the presence of many instruments, standard errors robust to the presence of many instruments and many exogenous regressors, and standard errors robust to the presence of direct effects of the instruments on the outcome.

The simulation results are reported in Table 5 for the case with valid instruments ($\Lambda_{11} = 0$). In the case with valid instruments, liml performs best, consistent with its efficiency properties. The mbtsls and btsls estimators do almost, but not quite as well. The bekker standard errors do well, the adjustment for many exogenous variables makes virtually no difference for coverage. The tsls estimator performs poorly, not surprising given the presence of many instruments.

When we simulate data with $\Lambda_{11} > 0$ and the instruments are not valid, the results change considerably. The liml estimator now performs very poorly. It has substantial bias and the coverage rates are low. Both the btsls and mbtsls estimators do well in terms of bias and median absolute deviation. Adjusting the variance for the presence of many exogenous covariates makes little difference, but the adjustment to allow for the presence of direct effects makes a considerable difference.

8 Conclusion

In this paper we analyze settings with many instruments where each separate instrument might have a direct effect on the outcome. We show that liml is particularly sensitive to such direct effects. In contrast, a modified version of the bias-corrected tsls estimator is robust to such direct effects if these direct effects are uncorrelated with the direct effects of the instrument on the endogenous

regressor. We argue in the context of some applications that this orthogonality condition has empirical content. In this setting the choice between liml and the mbtsls estimator depends on a trade-off between efficiency and robustness. In practice we recommend that researchers test for the presence of direct effects under the assumption of orthogonality of the direct effects, and that they compare liml and mbtsls estimates.

Appendices

We first define some additional notation. Write the reduced-form based on Equations (3.1) as:

$$\begin{pmatrix} Y_i & X_i \end{pmatrix} = Z_i' \begin{pmatrix} \psi_1 & \pi_1 \end{pmatrix} + W_i \begin{pmatrix} \psi_2 & \pi_2 \end{pmatrix} + V_i',$$

where $\psi_1 = \gamma + \pi_1\beta$ and $\psi_2 = \delta + \pi_2\beta$, and $V_i = (\epsilon_i + \nu_i\beta, \nu_i)'$, and let $\mathbf{V}$ be the $N \times 2$ matrix with ith row equal to V_i'. Denote the upper $K_N \times 2$ submatrix of the matrix of reduced-form coefficients by $\Pi = (\psi_1, \pi_1)$. Let

$$\overline{\mathbf{Y}}_\perp = \mathbf{M_W} \begin{pmatrix} \mathbf{Y} & \mathbf{X} \end{pmatrix}$$

denote the full matrix of endogenous variables after projecting out the exogenous covariates. Let:

$$\Gamma = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix}.$$

Let $\Omega = \mathbb{E}[V_i V_i']$ denote the reduced-form covariance matrix. Then:

$$\Omega = \Gamma^{-1\prime}\Sigma\Gamma^{-1} = \begin{pmatrix} \Sigma_{11} + 2\Sigma_{12}\beta + \Sigma_{22}\beta^2 & \Sigma_{12} + \Sigma_{22}\beta \\ \Sigma_{21} + \Sigma_{22}\beta & \Sigma_{22} \end{pmatrix}.$$

Let $\mathcal{W}_d(f, V, V^{-1}M)$ denote a d-dimensional non-central Wishart distribution with f degrees of freedom, scale parameter V, and non-centrality parameter M. Let $\mathbf{S}^{1/2}$ denote the symmetric square root of a symmetric positive semi-definite matrix $\mathbf{S}$.

Appendix A Auxilliary Lemmata

Lemma A.1 (Lemmata 1 and 2, Bekker, 1994)**.** *Consider the quadratic form $Q = (M+U)'C(M+U)$, where $M \in \mathbb{R}^{N\times S}, C \in \mathbb{R}^{N\times N}$ are non-stochastic, C is symmetric and idempotent with rank J_N which may depend on N, and $U = (u_1, \ldots, u_N)'$, with $u_i \sim [0, \Omega]$ iid. Let $a \in \mathbb{R}^S$ be a non-stochastic vector. Then:*

(i) *If u_i has finite fourth moments:*

$$\begin{aligned} \mathbb{E}[Q \mid C] &= M'CM + J_N\Omega, \\ \operatorname{var}(Qa \mid C) &= a'\Omega a M'CM + a'M'CMa\Omega + \Omega aa'M'CM + MCMaa'\Omega \\ &\quad + J_N(a'\Omega a\Omega + \Omega aa'\Omega) \\ &\quad + d_C'd_C[\mathbb{E}(a'u)^2uu' - a'\Omega aa'\Omega - a'\Omega a\Omega] + 2d_C'CMa\mathbb{E}[(a'u)uu'] \\ &\quad + M'Cd_C\mathbb{E}[(a'u)^2u'] + \mathbb{E}[(a'u)^2u]d_C'CM, \end{aligned}$$

where $d_C = \operatorname{diag}(C)$. If the distribution of u_i is Normal, the last two lines of the variance expression equals zero.

(ii) *Suppose that the distribution of u_i is Normal, and that, as $N \to \infty$:*

$$M'CM/N \to Q_{CM}, \qquad J_N/N \to \alpha_r,$$

where the elements c_{is} of C may depend on N. Then:

$$\sqrt{N}\left(Qa/N - \mathbb{E}Qa/N\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

where

$$V = a'\Omega a Q_{CM} + a'Q_{CM}a\Omega + \Omega aa'Q_{CM} + Q_{CM}aa'\Omega + \alpha_r(a'\Omega a\Omega + \Omega aa'\Omega).$$

Lemma A.2. *Consider a sequence of independent random matrices $\{X_N\}_{N=1}^{\infty}$ with distributions $X_N \sim \mathcal{W}_S(J_N, \Omega, \Omega^{-1}\Xi_N)$. Suppose that $\Xi_N/N \to \Xi$, and that $J_N/N = \alpha + o(N^{-1/2})$, $\alpha > 0$. Then, for any vector $a \in \mathbb{R}^S$*

$$N^{-1/2}\left(X_N a/N - (\Xi_N/N + \alpha\Omega)a\right) \xrightarrow{d} \mathcal{N}\left(0, (a'\Omega a\Xi + a'\Xi a\Omega + \Omega aa'\Xi + \Xi aa'\Omega) + \alpha(a'\Omega a\Omega + \Omega aa'\Omega)\right).$$

Proof. By definition of a non-central Wishart distribution, we can decompose $X_N = (U+M)'(U+M)$, where $U = (u_1, \ldots, u_{J_N})'$, $u_j \sim N(0,\Omega)$ iid, $M'M = \Xi_N$, and $\Xi_N/J_N \to \Xi/\alpha$. Hence, we can apply Lemma A.1 (ii) with $C = \mathbf{I}_{J_N}$ to get:

$$J_N^{-1/2}(X_N a - (\Xi_N + J_N\Omega)a) \xrightarrow{d} \mathcal{N}\left(0, \alpha^{-1}(a'\Omega a\Xi + a'\Xi a\Omega + \Omega aa'\Xi + \Xi aa'\Omega) + a'\Omega a\Omega + \Omega aa'\Omega\right),$$

which yields the result. □

Lemma A.3. *Suppose Assumptions 1, 2(i), 3 and 4 hold. Then:*

$$\overline{\mathbf{Y}}_{\perp}'\overline{\mathbf{Y}}_{\perp}/N \xrightarrow{p} \Psi + (1-\alpha_L)\Omega, \tag{A.1a}$$

$$\overline{\mathbf{Y}}_{\perp}'\mathbf{P}_{\mathbf{Z}_{\perp}}\overline{\mathbf{Y}}_{\perp}/N \xrightarrow{p} \Psi + \alpha_K\Omega, \tag{A.1b}$$

where

$$\Psi = \begin{pmatrix} \Lambda_{11} + 2\Lambda_{12}\beta + \Lambda_{22}\beta^2 & \Lambda_{12} + \Lambda_{22}\beta \\ \Lambda_{12} + \Lambda_{22}\beta & \Lambda_{22}. \end{pmatrix} \tag{A.2}$$

These probability limits also hold conditional on $\overline{\mathbf{Z}}$.

Proof. First we establish the probability limit of $\mathbf{V}'\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{V}/N$. By Lemma A.1 (i):

$$\mathbb{E}[\mathbf{V}'\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{V}/N \mid \mathbf{Z}_{\perp}] = (K_N/N)\Omega. \tag{A.3}$$

Fix $a \in \mathbb{R}^2$. Since $\mathbf{P}_{\mathbf{Z}_{\perp}}$ is a projection matrix, $0 \leq (\mathbf{P}_{\mathbf{Z}_{\perp}})_{ii} \leq 1$. Hence, $\sum_i (\mathbf{P}_{\mathbf{Z}_{\perp}})_{ii}^2 \leq \sum_i (\mathbf{P}_{\mathbf{Z}_{\perp}})_{ii} \leq$

K_N. Therefore:

$$\begin{aligned}
\operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N) &= \mathbb{E}\operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N \mid \mathbf{P}_{\mathbf{Z}_\perp}) \\
&= \mathbb{E}\left[\operatorname{tr}(\mathbf{P}_{\mathbf{Z}_\perp}/N^2)\right](a'\Omega a\Omega + \Omega aa'\Omega) \\
&\quad + \mathbb{E}\left[N^{-2}\textstyle\sum_i(\mathbf{P}_{\mathbf{Z}_\perp})_{ii}^2\right][\mathbb{E}(a'V_i)^2V_iV_i' - a'\Omega aa'\Omega - a'\Omega a\Omega] \\
&\leq \frac{K_N}{N^2}(a'\Omega a\Omega + \Omega aa'\Omega) + \frac{K_N}{N^2}[\mathbb{E}(a'v_i)^2v_iv_i' - a'\Omega aa'\Omega - a'\Omega a\Omega] \\
&= O(K_N/N^2).
\end{aligned} \tag{A.4}$$

Combining Equations (A.3) and (A.4) with Assumption 3 yields :

$$\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \xrightarrow{p} \alpha_K\Omega. \tag{A.5}$$

By similar arguments:

$$\mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \xrightarrow{p} (1-\alpha_L)\Omega. \tag{A.6}$$

Next, by Assumption 2 (i), $\mathbb{E}[\Pi'\mathbf{Z}_\perp'\mathbf{V}/N \mid \mathbf{Z}_\perp, \Pi] = 0$, so that:

$$\begin{aligned}
\operatorname{var}\left(\Pi'\mathbf{Z}_\perp'\mathbf{V}a/N\right) &= \mathbb{E}\left[\operatorname{var}\left(\Pi'\mathbf{Z}_\perp'\mathbf{V}a/N \mid \mathbf{Z}_\perp, \Pi\right)\right] = (a'\Omega a)\mathbb{E}\left[\Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi/N^2\right] \\
&= (a'\Omega a)\Gamma^{-1'}\mathbb{E}\left[\Lambda_N/N^2\right]\Gamma^{-1} = O(1/N),
\end{aligned}$$

where the last equality follows by Assumption 4. Consequently:

$$\Pi\mathbf{Z}_\perp'\mathbf{V}/N \xrightarrow{p} 0. \tag{A.7}$$

Combining the representation $\mathbf{Y}_\perp = \mathbf{Z}_\perp\Pi + \mathbf{V}_\perp$ with the limits in Equations (A.6) and (A.7), and Assumption 4 establishes (A.1a):

$$\begin{aligned}
\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N &= \Pi'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi/N + \Pi'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi/N + \mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \\
&= \Gamma^{-1'}\Lambda_N\Gamma^{-1}/N + (1-\alpha_L)\Omega + o_p(1) \\
&= \Psi + (1-\alpha_L)\Omega + o_p(1).
\end{aligned}$$

Claim (A.1b) follows by similar arguments from Equations (A.5) and (A.7):

$$\begin{aligned}
\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= \Pi'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi/N + \Pi'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi/N + \mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \\
&\xrightarrow{p} \Psi + \alpha_K\Omega.
\end{aligned}$$

This concludes the proof. □

Appendix B Proofs of Theorems

Proof of Theorem 1. Combining Lemma A.3 with the condition $\hat{\kappa} = \kappa + o_p(1)$ yields:

$$\begin{aligned}(1-\hat{\kappa})\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N + \hat{\kappa}\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= (1-\kappa)(\Psi + (1-\alpha_L)\Omega) + \kappa\cdot(\Psi + \alpha_K\Omega) + o_p(1) \\ &= \Psi + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)\Omega + o_p(1).\end{aligned} \tag{B.1}$$

Since $\Sigma_{22} = \Omega_{22}$, the (2,2) element of (B.1) is given by:

$$(1-\hat{\kappa})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{\kappa}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N = \Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)\Sigma_{22} + o_p(1).$$

By the condition on κ, $\Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)\Sigma_{22} > 0$, so that:

$$\left((1-\hat{\kappa})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{\kappa}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N\right)^{-1} = (\Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)\Sigma_{22})^{-1} + o_p(1). \tag{B.2}$$

The (1,2) element in Equation (B.1) is given by:

$$\begin{aligned}&(1-\hat{\kappa})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{\kappa}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp/N \\ &= \Lambda_{12} + \Lambda_{22}\beta + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)\Omega_{12} + o_p(1) \\ &\qquad = \Lambda_{12} + (1-\alpha_L - (1-\alpha_K-\alpha_L)\kappa)(\Sigma_{12} + \Sigma_{22}\beta) + \Lambda_{22}\beta + o_p(1).\end{aligned} \tag{B.3}$$

Combining Equations (B.2) and (B.3) with Slutsky's lemma then yields

$$\hat{\beta}_{\hat{\kappa}} = \frac{(1-\hat{\kappa})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{\kappa}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp/N}{(1-\hat{\kappa})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{\kappa}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N} = \beta + \frac{\Lambda_{12} + ((1-\kappa)(1-\alpha_L) + \alpha_K\kappa)\Sigma_{12}}{\Lambda_{22} + ((1-\kappa)(1-\alpha_L) + \alpha_K\kappa)\Sigma_{22}} + o_p(1).$$

□

Proof of Corollary 1. The results for tsls, btsls and mbtsls follow directly from Theorem 1. We therefore just need to derive the results for liml. Define

$$\hat{Q}_N(\phi) = \frac{\phi'\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N\phi}{\phi'\overline{\mathbf{Y}}_\perp'M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N\phi}.$$

Then

$$\hat{\kappa}_{\text{liml}} = \min_{\tilde{\beta}} \frac{(1,-\tilde{\beta})\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N(1,-\tilde{\beta})'}{(1,-\tilde{\beta})\overline{\mathbf{Y}}_\perp'M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N(1,-\tilde{\beta})'} = \min_{\phi\in S^1}\hat{Q}_N(\phi),$$

where S^1 denotes the unit circle in $\mathbb{R}^2$. Applying Lemma A.3 yields

$$\hat{Q}_N(\phi) \xrightarrow{p} \frac{\phi'(\Psi + (1-\alpha_L)\Omega)\phi}{(1-\alpha_L-\alpha_K)\phi'\Omega\phi} \equiv \frac{\phi'T\phi}{\phi'T_\perp\phi} \equiv Q(\phi),$$

where we define $T = \Psi + (1-\alpha_L)\Omega$ and $T_\perp = (1-\alpha_L-\alpha_K)\Omega$. Assumption 2 (i) guarantees that

the denominator is non-zero for any value of ϕ. The minimum of $Q(\phi)$ is achieved at

$$\begin{aligned}\min_{\phi\in S^1} Q(\phi) &= \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{1}{1-\alpha_L-\alpha_K}\min_{\phi\in S^1}\frac{\phi'\Psi\phi}{\phi'\Omega\phi} \\ &= \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\min \operatorname{eig}(\Sigma^{-1}\Lambda)}{1-\alpha_K-\alpha_L} = \kappa_{\text{liml}},\end{aligned}$$

where the last line follows since the eigenvalues of $\Omega^{-1}\Psi$ correspond to the eigenvalues of $\Sigma^{-1}\Lambda$. The minimand ϕ_{liml} is given by the eigenvector corresponding to the smallest eigenvalue of the matrix

$$\frac{1}{1-\alpha_K-\alpha_L}\Omega^{-1}\left(\Psi + (1-\alpha_L)\Omega\right).$$

We now need to show that

$$\hat{\kappa}_{\text{liml}} - \kappa_{\text{liml}} = \min_{\phi\in S^1}\hat{Q}_N(\phi) - Q(\phi_{\text{liml}}) \xrightarrow{p} 0. \tag{B.4}$$

To this end, we first show that the convergence of the objective function is uniform,

$$\sup_{\phi\in S^1}|\hat{Q}_N(\phi) - Q(\phi)| \xrightarrow{p} 0. \tag{B.5}$$

Fix $\phi \in S^1$. By the triangle inequality,

$$\begin{aligned}&|\hat{Q}_N(\phi) - Q(\phi)| \leq \\ &\leq \frac{1}{|\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp\phi/N|}\left|\phi'\overline{\mathbf{Y}}'_\perp\overline{\mathbf{Y}}_\perp\phi/N - Q(\phi)\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp\phi/N\right| \\ &= \frac{1}{|\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp\phi/N|}\left|\phi'(\overline{\mathbf{Y}}'_\perp\overline{\mathbf{Y}}_\perp/N - T)\phi - Q(\phi)\phi'\left(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp\right)\phi\right| \\ &\leq \frac{1}{|\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp\phi/N|}\left(\left|\phi'(\overline{\mathbf{Y}}'_\perp\overline{\mathbf{Y}}_\perp/N - T)\phi\right| + Q(\phi)\left|\phi'\left(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp\right)\phi\right|\right).\end{aligned} \tag{B.6}$$

We now need to bound all three terms in the expression uniformly in ϕ. Because the trace operator is the inner product under Frobenius norm, $\|\cdot\|_F$, by Cauchy-Schwarz inequality,

$$\begin{aligned}|\phi'(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp)\phi| &= \left|\operatorname{tr}\left((\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp)\phi\phi'\right)\right| \\ &\leq \sqrt{\operatorname{tr}((\phi\phi')^2)}\|(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp)\|_F \\ &= \|(\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N - T_\perp\|_F \\ &= o_p(1),\end{aligned}$$

where the third line follows since $\phi'\phi = 1$, and the last line follows since $\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N \xrightarrow{p} T_\perp$ by

Lemma A.3. By similar argument,

$$|\phi'(\overline{\mathbf{Y}}'_\perp \overline{\mathbf{Y}}_\perp/N - T)\phi| = o_p(1).$$

Finally, we bound the denominator. Because $\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp/N \xrightarrow{p} T_\perp > 0$, $\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp\phi/N > 0$ wpa1, so that wpa1 $1/(|\phi'\overline{\mathbf{Y}}'_\perp M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp\phi/N|) < C$ for some $C < \infty$. Applying these bounds and the fact that $Q(\phi)$ is bounded implies that the right-hand side in (B.6) is $o_p(1)$, which implies (B.5).

Next, denote the argmin of $\hat{Q}_N(\phi)$ by $\hat{\phi}$. Note that $\hat{\kappa}_{\text{liml}}$ and hence $\hat{\phi}$ exists wpa1. We can now establish (B.4), using the uniform convergence result (B.5),

$$\begin{aligned} Q(\phi_{\text{liml}}) \leq Q(\hat{\phi}) &= \hat{Q}_N(\hat{\phi}) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \leq \hat{Q}(\phi_{\text{liml}}) + (Q(\hat{\phi}) - \hat{Q}_n(\hat{\phi})) \\ &= Q(\phi_{\text{liml}}) + (\hat{Q}_N(\phi_{\text{liml}}) - Q(\phi_{\text{liml}})) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \\ &= Q(\phi_{\text{liml}}) + o_p(1). \end{aligned}$$

The probability limit for liml then follows by Theorem 1. □

Proof of Theorem 2. Under Assumption 2, we have:

$$\begin{aligned} \sqrt{\alpha_K}\begin{pmatrix} (\mathbf{Z}'_\perp\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}'_\perp\mathbf{Y} \\ (\mathbf{Z}'_\perp\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}'_\perp\mathbf{X} \end{pmatrix} \Big| \overline{\mathbf{Z}}, \tilde{\pi}_1, \tilde{\gamma} &\sim \mathcal{N}\left(\begin{pmatrix} \tilde{\pi}_1\beta + \tilde{\gamma} \\ \tilde{\pi}_1 \end{pmatrix}, \alpha_K\Omega \otimes \mathbf{I}_{K_N}\right), \\ \overline{\mathbf{Y}}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp \mid \overline{\mathbf{Z}}, \tilde{\pi}_1, \tilde{\gamma} &\sim \mathcal{W}_2(N - K_N - L_N, \Omega). \end{aligned}$$

Moreover, these two statistics are independent. Let $b = (1, -\beta)'$ and $a = (\beta, 1)$. Assumption 6 then implies that unconditionally,

$$\begin{aligned} \overline{\mathbf{Y}}'_\perp\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp &\sim \mathcal{W}_2(K_N, \Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega, \left(\Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega\right)^{-1} K_N\Gamma^{-1\prime}\mu\mu'\Gamma^{-1}/\alpha_K), \\ \overline{\mathbf{Y}}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp &\sim \mathcal{W}_2(N - K_N - L_N, \Omega), \end{aligned}$$

with the independence property preserved. Applying Lemma A.2 then after some algebra yields:

$$N^{1/2}\left(\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp b/N - (1 - \alpha_K - \alpha_L)\Sigma_{12}\right) \xrightarrow{d} \mathcal{N}\left(0, (1 - \alpha_K - \alpha_L)V_\Sigma\right), \tag{B.7a}$$

$$N^{1/2}\left(\mathbf{X}'_\perp\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/Nb - (\alpha_K\Sigma_{12})\right) \xrightarrow{d} \mathcal{N}\left(0, \alpha_K V_\Sigma + V_\Lambda\right), \tag{B.7b}$$

where

$$\begin{aligned} V_\Sigma &= \Sigma_{22}\Sigma_{11} + \Sigma_{12}^2, \\ V_\Lambda &= \Lambda_{22}\Sigma_{11} + \Lambda_{11}\Sigma_{22} + \alpha_K^{-1}\Lambda_{22}\Lambda_{11}. \end{aligned}$$

Equations (B.7) imply that

$$N^{1/2}\left(\mathbf{X}'_\perp\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N + (1 - \kappa_{\text{mbtsls}})\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N\right) b \xrightarrow{d} \mathcal{N}\left(0, V_\Lambda + \frac{\alpha_K(1 - \alpha_L)}{1 - \alpha_K - \alpha_L}V_\Sigma\right).$$

Since by Lemma A.3, $(\mathbf{X}'_\perp\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp N + (1 - \kappa_{\text{mbtsls}})\mathbf{X}'_\perp\mathbf{M}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N)^{-1} \xrightarrow{p} \Lambda_{22}^{-1} + o_p(1)$, this yields the claim in the theorem. □

References

Ackerberg, D. A. and Devereux, P. J. (2009). Improved Jive estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics*, **91** (2), 351–362.

Anatolyev, S. (2011). Instrumental variables estimation and inference in the presence of many exogenous regressors, unpublished manuscript.

— and Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, **27** (2), 427–441.

Anderson, T. W., Kunitomo, N. and Matsushita, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics*, **157** (2), 191–204.

— and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63.

Andrews, D. W. K., Moreira, M. J. and Stock, J. H. (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, **74** (3), 715–752.

Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67.

— and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014.

Ashley, R. (2009). Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics*, **24** (2), 325–337.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, **25** (1), 77–83.

Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, **62** (3), 657–681.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. B. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, (forthcoming).

Berkowitz, D., Caner, M. and Fang, Y. (2008). Are "Nearly Exogenous Instruments" reliable? *Economics Letters*, **101** (1), 20–23.

Caner, M. (2007). Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics, unpublished manuscript.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, **73** (5), 1673–1692.

—, —, Hausman, J. A., Newey, W. K. and Woutersen, T. (2012). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. *Econometric Theory*, **12** (1), 42–86.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? *Quarterly Journal of Economics*, **126** (4), 1593–1660.

Chioda, L. and Jansson, M. (2009). Optimal Invariant Inference When the Number of Instruments Is Large. *Econometric Theory*, **25** (3), 793–805.

Conley, T., Hansen, C. and Rossi, P. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, **94** (1), 260–272.

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, **9** (2), 222–240.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Donald, S. G. and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica*, **69** (5), 1161–1191.

Fisher, F. M. (1961). On the cost of approximate specification in simultaneous equation estimation. *Econometrica*, **29** (2), 139–170.

— (1966). The relative sensitivity to specification error of different k-class estimators. *Journal of the American Statistical Association*, **61** (314), 345–356.

— (1967). Approximate Specification and the Choice of a k-Class Estimator. *Journal of the American Statistical Association*, **62** (320), 1265–1276.

Flores, C. A. and Flores-Lagunes, A. (2010). Partial Identification of Local Average Treatment Effects with an Invalid Instrument, unpublished manuscript.

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets, unpublished manuscript.

Guggenberger, P. (2012). On the Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption. *Econometric Theory*, **28** (2), 387–421.

Hahn, J. and Hausman, J. A. (2005). IV Estimation with Valid and Invalid Instruments. *Annales d'Économie et de Statistique*, (79/80), 25–57.

Hansen, C. B., Hausman, J. A. and Newey, W. K. (2008). Estimation With Many Instrumental Variables. *Journal of Business and Economic Statistics*, **26** (4), 398–422.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, **3** (2), 211–255.

Kolesár, M. (2012). Random-Effects Approach to Inference With Many Instruments, unpublished manuscript.

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. and Imbens, G. W. (2011). Identification and Inference with Many Invalid Instruments, NBER working paper 17519.

Kraay, A. (2008). Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions, unpublished manuscript.

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association*, **75** (371), 693–700.

Mariano, R. S. (1973). Approximations to the Distribution Functions of Theil's k-Class Estimators. *Econometrica*, **41** (4), 715–721.

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica*, **51** (3), 821–841.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, **27** (4), 575–595.

Nevo, A. and Rosen, A. M. (2012). Identification with Imperfect Instruments. *Review of Economics and Statistics*, **94** (3), 659–671.

Reinhold, S. and Woutersen, T. (2011). Endogeneity and Imperfect Instruments in Applied Work: Deriving Bounds in a Semiparametric Model, unpublished manuscript.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of econometrics*, vol. 2, *Chapter 15*, Elsevier, pp. 881–935.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** (3), 393–415.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65** (3), 557–586.

Theil, H. (1961). *Economic Forecasts and Policy*. Amsterdam: Horth-Holland, 2nd edn.

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

van Hasselt, M. (2010). Many Instruments Asymptotic Approximations Under Nonnormal Error Distributions. *Econometric Theory*, **26** (02), 633–645.

Table 1: Estimates for Chetty *et al.* (2011) Data

		Standard Error			
Estimator	$\hat{\beta}$	classic	bekker	many exo	$\Lambda_{11} > 0$
Panel I: Grade 1 Test scores					
tsls	0.380	(0.038)			
liml	0.014	(0.047)	(0.052)	(0.052)	
btsls	0.221	(0.041)	(0.052)		
mbtsls	0.215	(0.041)	(0.052)	(0.052)	(0.066)
Panel II: Grade 2 Test scores					
tsls	0.389	(0.044)			
liml	0.108	(0.049)	(0.057)	(0.057)	
btsls	0.234	(0.047)	(0.059)		
mbtsls	0.226	(0.047)	(0.059)	(0.059)	(0.070)
Panel III: Grade 3 Test scores					
tsls	0.385	(0.048)			
liml	0.175	(0.052)	(0.061)	(0.061)	
btsls	0.238	(0.051)	(0.064)		
mbtsls	0.230	(0.051)	(0.064)	(0.064)	(0.070)

Notes: "classic" refers to conventional standard errors that assume fixed number of instruments, "bekker" refer to standard errors based on Bekker (1994) that allow for the number of instruments to increase with the sample size, "many exo" refers to standard errors also allow for many exogenous covariates, and "$\Lambda_{11} > 0$" are standard errors based on Theorem 2 that allow for direct effects of the instruments on outcome.

Table 2: Tests of Null Hypothesis $\Lambda_{11} = 0$ for Chetty *et al.* (2011) Data.

	Sargan		Adjusted Craig-Donald	
Outcome	Test Statistic	p-value	Test Statistic	p-value
Grade 1 Test scores	382.9	< 0.001	389.6	< 0.001
Grade 2 Test scores	319.3	< 0.001	320.8	< 0.001
Grade 3 Test scores	284.6	0.008	281.9	0.014

Table 3: Estimates for Angrist and Krueger (1991) Data ($N = 162,487$)

		Standard Error			
Estimator	$\hat{\beta}$	classic	bekker	many exo	$\Lambda_{11} > 0$
tsls	0.073	(0.018)			
liml	0.095	(0.018)	(0.040)	(0.040)	
btsls	0.097	(0.018)	(0.040)		
mbtsls	0.098	(0.018)	(0.040)	(0.040)	(0.040)

Notes: "classic" refers to conventional standard errors that assume fixed number of instruments, "bekker" refer to standard errors based on Bekker (1994) that allow for the number of instruments to increase with the sample size, "many exo" refers to standard errors also allow for many exogenous covariates, and "$\Lambda_{11} > 0$" are standard errors based on Theorem 2 that allow for direct effects of the instruments on outcome.

Table 4: Tests of Null Hypothesis $\Lambda_{11} = 0$ for Angrist and Krueger (1991) Data

	Sargan		Craig-Donald	
Outcome	Test Statistic	p-value	Test Statistic	p-value
log earnings	487.0	0.641	485.5	0.659

Table 5: Simulations: Median bias, Median absolute deviations, and coverage rates for nominal 95% confidence intervals for different estimators.

			Coverage			
Estimator	med. bias	MAD	classic	bekker	many exo	$\Lambda_{11} > 0$
			Panel I: $\Lambda_{11} = 0$			
tsls	0.147	0.023	2.7			
liml	0.001	0.032	90.9	95.0	95.0	
btsls	0.006	0.034	88.9	94.9		
mbtsls	0.000	0.035	88.6	94.9	94.9	95.6
			Panel II: $\Lambda_{11} = 0.7$			
tsls	0.148	0.031	7.6			
liml	−0.181	0.061	12.6	15.7	15.7	
btsls	0.005	0.045	76.5	86.4		
mbtsls	−0.001	0.046	76.2	86.3	86.3	94.0

Notes: "med. bias" and "MAD" stand for median bias and median absolute deviation. "classic" refers to conventional confidence intervals that assume fixed number of instruments, "bekker" refer to confidence intervals based on Bekker (1994) that allow for the number of instruments to increase with the sample size, "many exo" refers to confidence intervals that also allow for many exogenous covariates, and "$\Lambda_{11} > 0$" are confidence intervals based on Theorem 2 that allow for direct effects of the instruments on outcome.
The results are based on 10,000 simulation draws.