Finite Sample Bias from Instrumental Variables Analysis in Randomized Trials

Howard S. Bloom
Pei Zhu
(MDRC)

Fatih Unlu
(Abt Associates, Inc.)

August 2010

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

This paper was supported by funding from the Institute for Education Sciences (IES) of the U.S. Department of Education and the William T. Grant Foundation.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Ambrose Monell Foundation, The Annie E. Casey Foundation, Carnegie Corporation of New York, The Kresge Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

Acknowledgments

The authors thank Joshua Angrist, Hanley Chiang, Larry Hedges, Guido Imbens, David Judkins, Jeffrey Kling, Winston Lin, Sean Reardon, Jeffrey Smith, and Peter Schochet for their invaluable insights. The authors also thank seminar participants at Mathematica Policy Research, Inc., the Methods group meeting at IES, and participants at the 2009 SREE conference for their comments. The ideas presented and positions taken in the paper are solely the responsibility of the authors however, and do not necessarily reflect views of the funders.

For information about MDRC and copies of our publications, see our Web site: www.mdrc.org.

Abstract

This paper is the first step in a study of instrumental variables analysis with randomized trials to estimate the effects of settings on individuals. The goal of the study is to examine the strengths and weaknesses of the approach and present them in ways that are broadly accessible to applied quantitative social scientists. This paper begins with the methodological limitations of conventional ways to study causal relationships, such as cross-sectional regression, longitudinal regression, and latent variables analysis. It then examines finite sample bias for the simplest application of the alternative instrumental variable approach of a single-setting characteristic and individual outcome, and studies how "clustering" — when units of analysis (for example, students) are randomized or treated in groups (for example, by school) — affects instrument strength and finite sample bias. The last part of the paper extends the discussion to situations with multiple instruments for a single mediator and outcome where multiple instruments are constructed from information on treatment status for multiple studies, sites, or subgroups (aka ―strata). Specifically it addresses questions such as, "What happens when the treatment effect on the mediator is the same for all strata?" and "By how much must treatment effects on a mediator vary across strata in order for multiple instruments to reduce finite sample bias?" It also demonstrates that clustering does not affect answers to these questions.

Table of Contents

Appendixes

List of Tables and Figures

Table

Figure

Part 1

Introduction

This paper explores an alternative approach for studying the effects of characteristics of settings in which people live, work, study, or receive health care on their social, economic, academic, or health outcomes. The goal of the paper is to improve research methods for addressing questions like: What is the effect of:[1]

- neighborhood conditions on resident crime,
- organizational culture on employee turnover,
- classroom instruction on student achievement, or
- hospital practices on patient health?

The approach considered is instrumental variables analysis in randomized trials. This approach is gaining popularity and appears to have considerable potential for a broad range of applications. Unfortunately, it is subject to statistical problems that are not widely understood.

One of the most important problems is "finite sample bias," which as demonstrated by past research, can distort findings even from exceptionally large samples (Angrist and Krueger, 1991; Bound, Jaeger, and Baker, 1995). Unfortunately, the existing literature on this topic is highly technical and accessible mainly to econometricians and statisticians, even though the approach is potentially most valuable to applied social scientists. It is with this in mind that we attempt to unpack the problem in ways that promote a broader understanding of what produces it, how one can assess its severity and consequences, and therefore how one can decide when to use the new approach.

Given the opaqueness of the existing literature (at least to us), we derive from "first principles" each feature of finite sample bias that is discussed. Doing so helped us to develop a conceptual understanding of these features and hopefully will do so for others. Consequently, we include our derivations in appendixes for interested readers. In some cases, we re-derive findings that have been established previously. In other cases we derive findings that are new (we think). But in all cases, we try to provide derivations that facilitate a practical understanding of how to assess and address finite sample bias. Toward this end, we have tried to make each appendix self-contained, which in some cases has required repeating points that are discussed in the main body of the paper and in other appendixes.

[1] This paper grew out of the authors' attempt to use the new approach to study effects of instructional practices on student achievement.

The paper consists of three parts. Part 1 identifies the research question that motivated our interest in the new approach, illustrates limitations of conventional methods for addressing the question, describes the new approach for doing so, and introduces finite sample bias. This provides a conceptual framework for the paper.

Part 2 works through key features of finite sample bias for the simplest possible situation: analyses of a single setting feature based on a single instrumental variable (explained later). This makes it possible to illustrate the issues at play with a minimum of complexity. To simplify further, the discussion begins by considering samples that do not have individuals clustered within aggregate entities (for example, students are not clustered within specific schools). These findings are then generalized to samples with clustering, which better represent situations to which the new method is likely to apply. Throughout, the discussion uses simple graphs, algebra, and numerical examples to consider the factors that cause finite sample bias, how one can assess the severity of this problem, and how one can reduce its severity.

Part 3 of the paper generalizes findings to a more realistic situation: analyses of a single-setting feature based on multiple instrumental variables (explained later). This discussion, which is similar in style to Part 2, provides a step-by-step examination of the factors involved in finite sample bias, their causes, their symptoms, and their consequences.

Part 4 concludes the paper by briefly noting the next steps in our research on using instrumental variables analysis with randomized trials to study causal relationships. Among other things, this future work will attempt to generalize findings to analyses of multiple setting features based on multiple instrumental variables. This situation complicates the issues discussed below and raises important new ones.

Research Question

Our research is motivated by the question: How can one obtain valid and precise estimates of *causal* relationships between features of settings and individual outcomes? Specific examples of this general question exist throughout the applied social sciences. For example, education research has a strong interest in how different types of classroom instruction or features of teacher/student relationships affect student achievement.[2] Similarly, neighborhood research has had a long-standing interest in how levels of poverty, crime, segregation, or violence affect resident behavior.[3] Likewise, organizational research traditionally has focused on how factors like group culture affect productivity. In each case, the question of interest is about *cause and effect*, with a setting feature as the cause (or causal agent) and an individual outcome as the effect (or causal effect).

When addressing such questions, it is well known that "correlation does not necessarily mean causation." However, this point is often ignored as researchers and decision-makers grasp for badly needed answers to important questions. Recently however, there has been a marked increase in the rigor with which causal questions are being addressed, aided, and abetted by an increased use of randomized trials and rigorous quasi-experiments like regression discontinuity designs.[4] Furthermore, statisticians have begun to develop a systematic theory of causal inference.[5]

At the heart of this theory is the concept of *potential outcomes* under different states of the world with respect to an identifiable causal agent. Consider for example, the effect of teacher responsiveness on student achievement. The causal agent — teacher responsiveness — is a feature of educational settings, and the causal effect — student achievement — is an individual outcome. The causal relationship between these two constructs represents how different levels of the causal agent relate to different levels of the causal effect. In theory, this relationship exists for every individual in a target population. In other words, there is a specific individual outcome for each potential level of the causal agent. Consequently, a target population of interest has a distribution of potential outcomes for each level of the causal agent. The primary task for researchers is to estimate parameters of these distributions, especially their means.

Because our work grows out of evaluation research, we tend to view features of settings as intermediate outcomes through which interventions produce effects. Thus we typically conceptualize setting features as *mediators* of intervention effects. For example, an educational intervention might focus on

[2] Using conventional cross-sectional analyses, Gamse et al. (2008) and Jackson et al. (2007) examine associations between classroom instructional practices and student achievement in the studies of the federal Reading First and Early Reading First programs, respectively. Pianta et al. (2002) studies the relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes.

[3] See Ludwig and Kling (2007) for examples of studies that investigate such relationships.

[4] Since its initiation in 2002, the Institute of Education Sciences has funded 21 large-scale impact evaluations that employ randomized trials or regression discontinuity designs (Institute of Education Sciences, 2008).

[5] Donald Rubin has a series of publications on the counterfactual or potential outcome model for causal inference (for example, 1974, 1977, and 1978). Morgan and Winship (2007) provides a recent formulation of the topic.

improving teacher responsiveness in order to increase student achievement. In this case, teacher responsiveness is the mediator (M) of the causal effect of an intervention or "treatment" (T) on an individual outcome (Y).

Limitations of Conventional Approaches

As a point of departure, we consider first limitations of the most commonly used ways that evaluation studies estimate causal relationships between mediators and outcomes in general and between features of settings and individual outcomes in particular.

Cross-Sectional Regression Analysis

The simplest and most popular approach is cross-sectional regression analysis. Figure 1 presents a simplified graphical model of this approach. It specifies a causal effect $(\pi)$of a treatment (T) on a setting feature or mediator (M), which in turn, has a causal effect (β_{ca}) on an individual outcome (Y). The model also includes factors that affect the individual outcome and are correlated with the mediator. One of these factors (X_1) is observed and therefore can be controlled for statistically. The other factors (X_2 … X_q) are unobserved and cannot be controlled for statistically. Measures for all variables, except treatment status, contain random error.

Equation 1 below is a cross-sectional regression equation for this model. It typically is estimated using ordinary least squares (OLS) from data for a treatment group and control group.

$$Y_i = \alpha_{cs} + \beta_{cs}M_i + \beta_{cs1}X1_i + \nu_i \qquad (1)$$

where:

$Y_i, M_{i,}\ and\ X1_i$ = values of the outcome, mediator, and covariate for individual i,

$\alpha_{cs}, \beta_{cs}\ and\ \beta_{cs1}$ = a cross-sectional intercept and cross-sectional regression coefficients for the mediator and covariate,

ν_i = a random error term that is independently and identically distributed.

Estimates of the cross-sectional regression coefficient for the mediator, $\hat{\beta}_{cs}$, are used to measure the causal effect of the mediator on the outcome. There are three problems with doing so: omitted-variables bias, attenuation bias, and simultaneity bias. The first of these problems is well known, whereas the other two are less widely recognized. To simplify their discussion, we address each problem in the absence of the others, although in practice they tend to occur together.

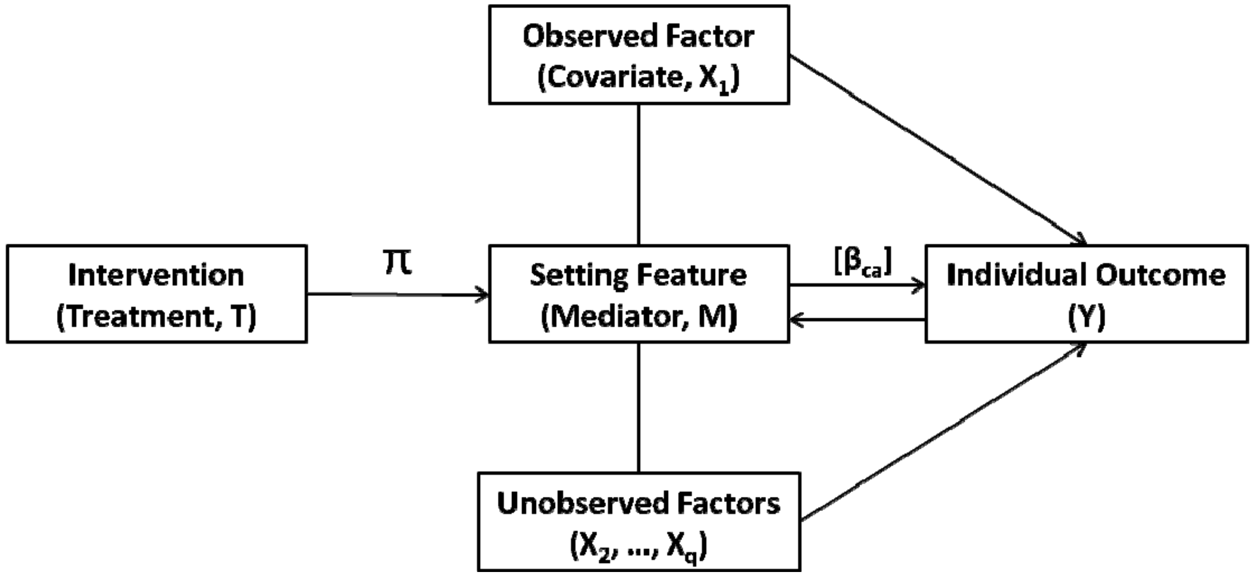
Omitted Variables Bias

Omitted variables bias is produced by correlations of unobserved factors (X_2 … X_q) with the mediator and outcome. These correlations provide an indirect path from M to Y, which produces a spurious

Figure 1

Path Model of Treatment, Mediator, Covariates, and Outcome for a Randomized Trial

(Given the Exclusion Restriction)



NOTE: Lines without arrows are used to indicate cross-sectional associations and arrows indicate the direction of causal effects.

(noncausal) observed association between them. The resulting estimates of the cross-sectional regression coefficient, $\hat{\beta}_{cs}$, include this spurious association as a bias. If correlations of the unobserved factors with the mediator and outcome have the same sign, they induce a positive bias in $\hat{\beta}_{cs}$; if they have opposite signs they induce a negative bias. Hence, the direction of this bias is often not predictable. What is predictable, however, is that cross-sectional regressions can produce biased estimates of the causal effect of the mediator on the individual outcome due to omitted variables.

Attenuation Bias

Attenuation bias is produced by random measurement error in the mediator, which can be especially problematic for features of settings, because they are difficult to measure reliably.[6] Appendix A demonstrates that absent omitted variables, simultaneity, or covariates (the simplest possible situation), the probability limit of $\hat{\beta}_{cs}$ is:

$$plim(\hat{\beta}_{OLS}) = \lambda\beta_{ca} \qquad (2)$$

[6] Economists often refer to bias from random measurement error in an independent variable as the problem of "errors-in-variables."

where λ is the reliability of the measured mediator and β_{ca} is the causal effect of the mediator on the outcome. The reliability of a measured mediator is the proportion of its total observed variation that is due to variation in true values, or: [7]

$$\lambda \equiv \frac{VAR(M_{true})}{VAR(M_{true})+VAR(M_{error})} \tag{3}$$

Consider the implications of Equation 2. If the measured mediator has a reliability of 0.80, then estimates of its regression coefficient will tend toward 80 percent of its causal effect on the outcome. If reliability is only 0.50 these estimates will tend toward 50 percent of the causal effect. Hence, measurement error attenuates coefficient estimates, potentially by a lot.

Appendix A demonstrates that the situation is even worse when a covariate, X_1, is included in the regression analysis. Specifically:[8]

$$plim(\beta_{OLS}) = [\frac{\lambda - R^2_{MX1}}{1-R^2_{MX1}}]\beta_{ca} \tag{4}$$

where $R^2_{MX1}$is the square of the correlation (R-squared) between the mediator and covariate. Table 1 illustrates the implications of this expression. It indicates that when R-squared is low, the attenuation proportion approximates the reliability of the mediator (which is the case without a covariate). However as R-squared increases, attenuation becomes more severe, and estimates of the mediator's regression coefficient tend toward a decreasing proportion of its causal effect on the individual outcome.[9]

Recent research suggests that the reliability of setting-level measures is much lower than was thought previously (Raudenbush et al., 2008). This is because recent research uses generalizability theory to consider all known sources of measurement error and their interactions together, whereas past research typically uses classical measurement theory, which considers only a single source of measurement error at a time.[10] For example, to assess a classroom observational protocol, classical measurement theory would consider error due to rater differences (inter-rater reliability), or item differences (inter-item consistency), or temporal differences (test-retest reliability) separately. In contrast, generalizability theory would consider all of these error components and their interactions together and report a composite measure of their cumulative effect on reliability. It is not surprising then that broader and more realistic measures

[7] Guilford (1965), pp. 439-440.

[8] To simplify the present discussion, we assume that the covariate X is measured without error. In future work, we will consider a covariate that is measured with random error.

[9] The increased attenuation produced by the covariate occurs because, as its correlation with the measured mediator increases, the covariate becomes an increasingly good "proxy" for the true mediator. Consequently, the estimated regression coefficient for the covariate, $\widehat{\beta}_{cs1}$, "absorbs" some of the causal relationship between the true mediator and the outcome. Simulations by the authors indicate that this phenomenon can be pronounced.

[10] Generalizability theory was introduced by Cronbach et al. (1972). Shavelson and Webb (1991) provide an excellent introduction to the topic and Brennan (2001) provides a comprehensive treatment of it.

Table 1

Attenuation Bias with a Covariate

Reliability of Mediator (λ)	**R-Square Between Mediator and Covariate (R^2_{MX1})**			
	0.2	**0.4**	**0.6**	**0.8**
0.8	0.75	0.67	0.5	0
0.6	0.50	0.33	0	
0.4	0.25	0		
0.2	0			

NOTES: ***(a)*** Figures in the table represent the probability limit of the estimated regression coefficient for the mediator ($\widehat{\beta_{cs}}$) as a proportion of the true causal effect of the mediator on the outcome (β_{ca}).
(b) Missing cells represent combinations of reliability and correlation that are not possible because R^2_{MX1} cannot exceed λ.

based on generalizability theory report lower reliability than do narrower and less realistic ones based on classical measurement theory.

The lower estimates of reliability for setting-level measures produced by assessments based on generalizability theory imply that attenuation bias is likely to be substantial for conventional cross-sectional regression analyses of relationships between features of settings and individual outcomes.

Simultaneity Bias

Simultaneity bias occurs in cross-sectional data when there is reciprocal causality between a mediator and an individual outcome. Figure 1 indicates this phenomenon with a two-way arrow between M and Y. This implies that changing the value of M causes the value of Y to change (which is the causal effect of interest), and changing the value of Y causes the value of M to change (which is the source of simultaneity bias). For example, consider a situation in which: (1) increased teacher responsiveness causes higher student achievement and (2) higher student achievement causes increased teacher responsiveness. A simple cross-sectional regression coefficient will reflect the composite effect of both of these causal paths. If the causal paths have the same signs, the regression coefficient will overstate the magnitude of the causal effect of M on Y. This means that the regression coefficient will overstate what would happen to Y if an exogenous change were induced in M. If the causal paths have opposite signs, the regression coefficient will understate the magnitude of the causal effect of interest. This means that the regression coefficient will understate what would happen to Y if an exogenous change were induced in M. Without knowing which of the preceding two cases exists, it is not possible to predict the direction of simultaneity bias.

Combined Bias

It is even more difficult to predict the combined effects of omitted variables bias, attenuation bias, and simultaneity bias. However, given the strong potential for attenuation bias, we believe that cross-sectional regressions often understate (potentially by a lot) the strength of causal relationships between setting features and individual outcomes. This might help to explain why it has been so difficult for researchers to observe these relationships empirically.

Longitudinal Regression Analysis

A second commonly used approach for studying the causal effects of mediators on outcomes is longitudinal regression analysis. These analyses regress observed *changes* in an individual outcome, ΔY_i, on observed changes in a mediator, ΔM_i, and a covariate, $\Delta X1_i$, yielding:

$$\Delta Y_i = \alpha_{lg} + \beta_{lg}\Delta M_i + \beta_{lg1}\Delta X1_i + \Delta \nu_i \qquad (5)$$

The primary benefit of this approach is that it eliminates bias due to omitted variables that do not change over time. This is because observed changes net them out. For example, if student motivation remains constant over time, it is possible to net out its influence through a regression of changes on changes. On the other hand, motivation — and other variables that are correlated with both M and Y — might change over time and thereby produce omitted variables bias in longitudinal analyses.

In addition, behavioral patterns that produce reciprocal causality in cross-sectional data most likely remain in longitudinal data.[11] So this approach is not likely to solve the problem of simultaneity bias. Furthermore, and perhaps most problematic, is that the reliability of change measures is typically much lower than that of their cross-sectional counterparts. Thus, attenuation bias is probably more severe for conventional longitudinal regression analyses than it is for conventional cross-sectional regression analyses (Deaton, 1997).

Latent Variables Analysis

One way to address attenuation bias in a causal analysis is to account directly for the reliability of the measured mediator. This approach is often referred to as latent variables analysis because it explicitly distinguishes between an observed measure and its underlying construct or latent variable. Several versions of the approach exist. One version uses external information about reliability from past research. This information can be imbedded in a regression analysis to account for measurement error in a mediator (Raudenbush, 2007). The information also can be used to adjust (after the fact) an estimated regression coefficient (using Equation 2 or 4). Another version of the approach is to design data collection for a

[11] Researchers sometimes try to reduce simultaneity bias using longitudinal data by specifying models with a lagged value of the mediator as an independent variable and current values of the individual outcome as the dependent variable. However, if there is serial correlation in the mediator and outcome (which is often likely), simultaneity bias remains a problem in such analyses.

given causal study that enables estimation of the reliability of its measured mediator. This approach can combine a statistical model of measurement with a statistical model of the causal relationship of interest and estimate them together. However, latent variables approaches per se do not deal with bias from omitted variables or simultaneity.

Path Analysis

Path analysis is used frequently for mediational analyses. It attempts to simultaneously model all relevant links in a causal chain or system of causal chains. A high-quality path analysis is carefully guided by preexisting theory that specifies how each variable in a model is related to the others. In this regard, path analysis represents an improvement over standard regression analysis. However, parameter estimates for path models are usually subject to the same sources of biases that plague regression analysis: attenuation, omitted variables, and simultaneity. Attenuation bias will exist unless information about the reliability of key independent variables is available. Omitted variables bias is always a risk, because the theories that underlie path analyses are typically quite limited. Simultaneity bias is often a risk unless there a source of exogenous variation (an instrument).

Bottom Line

For the preceding reasons, we do not believe that conventional approaches for estimating causal relationships between features of settings (mediators in intervention studies) and outcomes for individuals (final outcomes in intervention studies) are adequate for the task. Because of this, we have begun to explore alternative approaches in general and instrumental variables analysis with randomized trials in particular.

An Alternative Approach

Figure 1 illustrates the core elements of instrumental variables analysis in a randomized trial. Its logic is as follows. Treatment, T, has a *constant* causal effect, π, on mediator, M, which in turn, has a *constant* causal effect, β_{ca}, on outcome, Y. <u>If all of the treatment effect on the outcome is through the mediator</u> (an assumption that will be examined in our future work), the effect of treatment on the outcome equals $\pi\beta_{ca}$.[12] In a properly implemented randomized trial, the observed difference between mean values of the mediator for the treatment group and control group ($\Delta M/\Delta T$) is a consistent and unbiased estimate of the constant causal effect of treatment on the mediator, π. Furthermore, the observed treatment and control group difference in mean values for the outcome ($\Delta Y/\Delta T$) is a consistent and unbiased estimate of the constant causal effect of treatment on the outcome $\pi\beta_{ca}$. Consequently the ratio of these two estimators is a consistent estimator of the causal effect of the mediator on the outcome, or

[12] This fundamental condition is usually referred to as the “exclusion restriction.”

$$\text{plim}(\frac{\Delta Y}{\Delta M}) = \frac{\text{plim}(\frac{\Delta Y}{\Delta T})}{\text{plim}(\frac{\Delta M}{\Delta T})} = \frac{\pi\beta_{ca}}{\pi} = \beta_{ca} \qquad (6)$$

Equation 6 represents the probability limit of a Wald Estimator, which is the simplest form of instrumental variables analysis.[13] For example, if a randomized trial found that a particular form of professional development for teachers increased their average responsiveness to students by 10 points on scale A ($\frac{\Delta M}{\Delta T} = 10$) and found that average student achievement increased by 20 points on scale B ($\frac{\Delta Y}{\Delta T} = 20$) a consistent estimate of the causal effect of teacher responsiveness on student achievement is

$$\frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{20}{10} = 2\ scale\ B\ points\ per\ scale\ A\ point \qquad (7)$$

Figure 2 illustrates this analysis graphically, with values of the mediator, M, on the horizontal axis and values of the outcome, Y, on the vertical axis. The circle in the figure represents mean values of M and Y for the comparison group; the square represents mean values of M and Y for the treatment group; the origin of the graph represents mean values of M and Y for the combined study sample.

The horizontal distance between the circle and the square represents the treatment group and control group difference in the mean value of the mediator ($\frac{\Delta M}{\Delta T}$), which is the estimated effect of treatment on the mediator. The corresponding vertical distance represents the treatment group and control difference in the mean value of the individual outcome ($\frac{\Delta Y}{\Delta T}$), which is the estimated effect of treatment on the outcome. The slope of the line between the circle and square equals the ratio of the two estimators ($\frac{\Delta Y}{\Delta M}$), which is the estimated effect of the mediator on the outcome. This slope is the graphical counterpart of a Wald Estimator.

To develop a more general framework for such analyses, we specify it as the following two-stage least-squares (TSLS) regression model.[14]

First stage:

$$M_i = \mu + \pi T_i + \varepsilon_i \qquad (8)$$

Second stage:

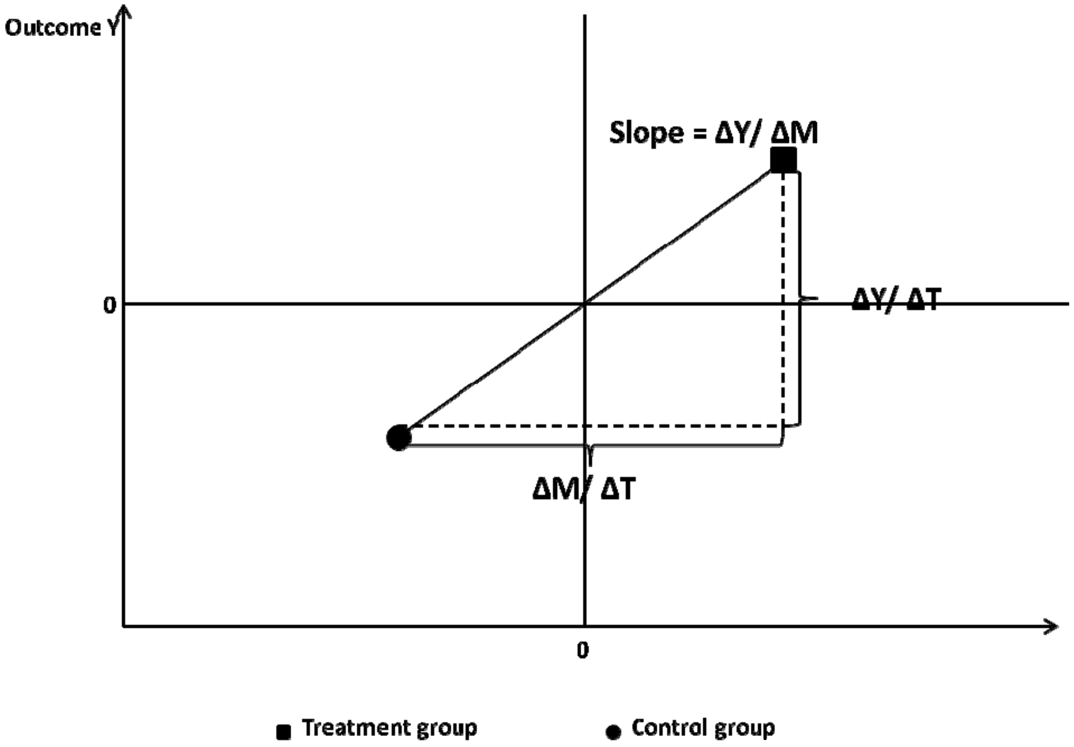
$$Y_i = \alpha + \beta_{ca} M_i + \nu_i \qquad (9)$$

[13] The Wald estimator was first discussed by Wald (1940).

[14] We refer to two-stage least-squares estimation as TSLS, which should not be confused with three-stage least-squares estimation. Another method for estimating instrumental variables models is limited information maximum likelihood or LIML. For a single mediator and instrument LIML is the same as TSLS. But with multiple instruments the two estimation procedures differ. We will explore the properties of LIML in future work.

Figure 2

Graphical Analysis for a Single Mediator and Instrument



The first-stage regression represents the relationship between treatment status and the mediator. OLS estimates of its intercept and coefficient are used to predict the value of the mediator for each sample member. These predicted values, $\widehat{M}_i$, are substituted for actual values of M_i, in the second-stage regression, whose intercept and coefficients are estimated using OLS, adjusting standard errors to account for using predicted instead of actual values of M.[15] The resulting estimate of the second-stage regression coefficient, $\hat{\beta}_{TSLS}$, is identical to a Wald Estimator.[16]

The intuition behind two-stage least-squares analysis is as follows. Omitted-variables, measurement error, and reciprocal causality "contaminate" the variation of mediator values across sample members. This contaminated or endogenous variation produces spurious correlation between observed values of M and Y. (They induce a correlation between M and the error term, ν, in Equation 9). This spurious correlation causes the bias in conventional estimators of the causal relationship between M and Y.

[15] See Greene (1997, p. 295, p. 742) for a discussion of this adjustment to the standard errors.
[16] Angrist and Krueger (1999).

Instrumental variables analysis *operates like surgery*. It attempts to remove all contaminated (endogenous) variation in M while leaving as much systematic uncontaminated (exogenous) variation as possible. Intuitively, the procedure works as follows. Randomized treatment status can induce systematic uncontaminated variation in the mediator, as long as all of the effect of treatment status on the individual outcome is *through* its effect on the mediator. In this way, treatment status is an "*instrument* of uncontaminated change" in the mediator.

Assume for a moment that we know the treatment effect on the mediator, π, and we use it to predict differences in mediator values for treatment and control group members. These *treatment-induced differences* are the only source of variation in predicted mediator values, and the larger the treatment effect is, the more treatment-induced variation there is across sample members. If treatment status affects individual outcomes only through the mediator, then treatment-induced variation in the mediator (*tiv*) is uncontaminated or exogenous to variation in individual outcomes. (It is not correlated with ν in Equation 9.) Consequently, treatment-induced variation does not carry with it a spurious correlation between M and Y. Using predicted values in the second-stage regression therefore provides consistent estimates of the causal effect of the mediator on the outcome.

Finite Sample Bias with the Alternative Approach

With an infinite sample, OLS analysis of the first-stage regression (Equation 8) provides perfect knowledge of the true effect of treatment on the mediator, π. This, in turn, makes it possible to compute the exact treatment-induced variation in mediator values. Using this information in a second-stage regression produces an unbiased and consistent estimate of the causal effect of the mediator on the individual outcome.

With a finite sample — which is all that exists in practice — random assignment can produce unbiased and consistent *estimates* ($\hat{\pi}$) of treatment effects on mediators, but it cannot produce perfect knowledge of this treatment effect. Error in these estimates produces error-induced variation (*eiv*) in predicted values of the mediator that is contaminated (endogenous to Y). Therefore in finite samples, instrumental variables analyses produce estimates of the causal effect of a mediator that reflect both treatment-induced variation, which is uncontaminated, and error-induced variation, which is contaminated.

As will be demonstrated, the central tendency (mean or median) of the sampling distribution of an instrumental variables estimator of the causal effect of a mediator on an individual outcome approximates an average of the true causal effect, $\beta_{ca,}$ and the underlying cross-sectional relationship, β_{cs}, with weights proportional to the expected values TIV and EIV of *tiv* and *eiv*, respectively. In symbols:

$$CENTRALTENDENCY(\hat{\beta}_{TSLS}) \approx \left[\frac{TIV}{TIV+EIV}\right]\beta_{ca} + [\frac{EIV}{TIV+EIV}]\beta_{cs} \qquad (10)$$

Equation 10 is the key to understanding the causes, consequences, and potential solutions to finite sample bias.[17]

[17] For reasons discussed in the paper, we will focus on the median of this estimator rather than its mean or expected value.

Part 2

Finite Sample Bias with a Single Mediator and a Single Randomized Instrument

This part of the paper examines finite sample bias for a single mediator and a single instrument. It first develops a conceptual model of the problem, then examines the problem for samples that are not clustered, and lastly generalizes findings to clustered samples.

Our findings apply to two different prototypical situations, which are referred to interchangeably. The first situation is an ***individual-level analysis***, where the mediator and outcome vary across individuals. For example, the mediator might be individual student engagement and the outcome might be individual student achievement. Students might be clustered — for example, by school or by classroom — or they might not be clustered — for example, if each were from a different school. In either case, the unit of analysis is the individual student. If individuals are clustered — either because they are randomized in groups or because they are treated in groups — then the instrument (treatment status) varies only by cluster. If individuals are not clustered (because they are randomized *and* treated independently) the instrument varies by individual.

The second situation, which is the substantive motivation for the present paper, is a ***setting-level analysis***. In this case, the mediator is a setting-level characteristic, and the outcome is either inherently a setting-level characteristic or is an individual-level characteristic that is aggregated to the setting level — usually by averaging. For example, the setting mediator might be a specific classroom instructional practice, and the setting outcome might be average student achievement for each classroom. Settings (for example, classrooms) might themselves be clustered within higher-level aggregates (for example, schools) or they might not be. But in either case, the setting is the unit of analysis. If settings are randomized and treated independently, then the instrument (treatment status) varies by setting. If settings are randomized or treated by cluster, then the instrument varies by cluster.

This aggregate framework for studying the effects of setting-level mediators is used because it represents how instrumental variables estimation usually is applied to them. Although the framework does not explicitly specify the clustering of individuals (for example, students) within settings (for example, classrooms) through a multilevel model, it provides unbiased estimates of relevant parameters and properly accounts for "within-setting" clustering when estimating standard errors. This convenient result occurs because values of the mediator do not vary across individuals within a setting, and the observed variation in setting mean values of the individual outcome properly reflects its between-setting and within-setting variance components.

The convention we use for both individual-level situations and setting-level situations is as follows: We refer to settings or individuals as ***units***, and we refer to groups of units that are randomized and/or treated together as ***clusters***.

Conceptual Model

Figure 3 reframes our situation in a way that makes it easier to *see* how finite sample bias works. The top line in the figure represents causal relationships between a treatment, a mediator, and an outcome. The bottom line represents counterfactual values of the mediator and outcome (M_* and Y_*). These are values that would occur in the absence of treatment. Although counterfactual values exist for all sample members, they can be observed only for control group members. The figure indicates the following cross-sectional relationship between M_* and Y_*:

$$Y_{*i} = \alpha_{cs} + \beta_{cs} M_{*i} + \upsilon_i \tag{11}$$

The cross-sectional coefficient, β_{cs}, in Equation 11 reflects: (1) the true causal effect of M_* on Y_*, (2) omitted variables bias from unobserved variables that affect both M_* and Y_*, (3) simultaneity bias from a reverse causal effect of Y_* on M_*, and (4) attenuation bias from measurement error in M_*.

Consider how the cross-sectional relationship in Equation 11 affects TSLS or Wald estimates of the causal coefficient, β_{ca}. First note that the predicted value of a mediator from an estimated first-stage regression is:

$$\widehat{M}_i = \hat{\mu} + \hat{\pi} T_i \tag{12}$$

The intercept and coefficient of this regression are estimated with errors, ε_π and ε_μ, such that:

$$\hat{\pi} = \pi + \varepsilon_\pi \tag{13}$$

$$\hat{\mu} = \mu + \varepsilon_\mu \tag{14}$$

Substituting Equations 13 and 14 into Equation 12 yields:

$$\widehat{M}_i = (\mu + \varepsilon_\mu) + (\pi + \varepsilon_\pi) T_i \tag{15}$$

Equation 15 indicates that predicted values of the mediator reflect true values of the intercept and slope for the first-stage regression (μ and π) plus their estimation errors (ε_μ and ε_π). Of primary concern is estimation error, ε_π, for the slope, which reflects the treatment/control group “mismatch” on counterfactual values of the mediator — or *mismatch error* for short. Specifically:

$$\varepsilon_\pi = \overline{M}_{*T} - \overline{M}_{*C} \tag{16}$$

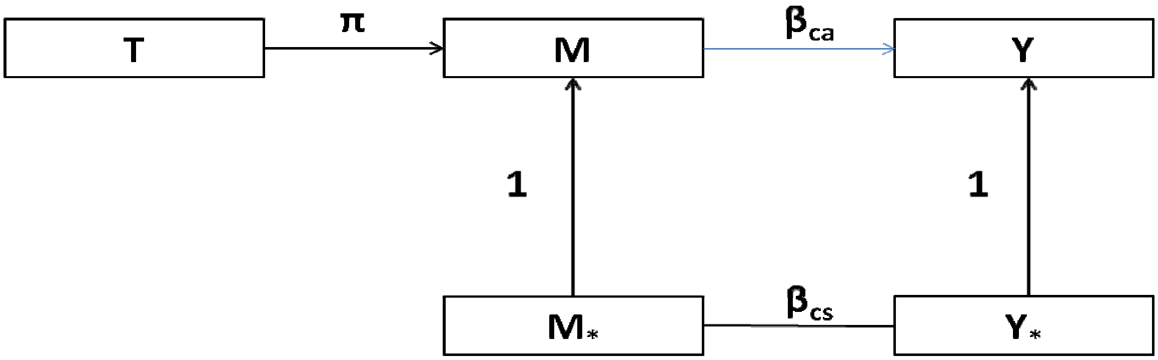
If ε_π is positive, the mean value of M_* is higher for the treatment group than for the control group. If ε_π is negative, the mean value of M_* is lower for the treatment group than for control group. Either result is equally likely to occur by chance, given the “luck of the draw” from randomization.

Substituting Equation 16 into Equation 15 and rearranging terms yields:

$$\widehat{M}_i = (\mu + \varepsilon_\mu) + (\pi) T_i + (\overline{M}_{*T} - \overline{M}_{*C}) T_i \tag{17}$$

Figure 3

Path Model of a Single Mediator and Instrument with a Randomized Trial (Given the Exclusion Restriction)



NOTE: Lines without arrows are used to indicate cross-sectional associations, and arrows indicate the direction of causal effects.

Equation 17 indicates that predicted values of the mediator, $\widehat{M}_i$, are a linear function of the treatment effect on the mediator, π, and mismatch error $(\bar{M}_{*T} - \bar{M}_{*C})$. Variation in $\widehat{M}_i$ due to the treatment effect is treatment-induced, whereas variation due to mismatch error is error-induced.

To express the preceding argument formally, consider the sum of squares of the predicted mediator which can be expressed as the following:

$$\begin{aligned}\sum_{i=1}^{N}(\widehat{M_\iota} - \bar{\widehat{M}})^2 &= \sum_{i=1}^{N}\{[\hat{\mu} + (\pi + \epsilon_\pi)T_i] - [\hat{\mu} + (\pi + \varepsilon_\pi)\bar{T}]\}^2 \\ &= \sum_{i=1}^{N}[(\pi + \varepsilon_\pi)(T_i - \bar{T})]^2 \\ &= (\pi + \varepsilon_\pi)^2 \sum_{i=1}^{N}[(T_i - \bar{T})]^2 \\ &= \underbrace{\pi^2 \sum_{i=1}^{N}[(T_i - \bar{T})]^2}_{tiv} + \underbrace{\varepsilon_\pi^2 \sum_{i=1}^{N}[(T_i - \bar{T})]^2}_{eiv} + 2\pi\varepsilon_\pi \sum_{i=1}^{N}[(T_i - \bar{T})]^2 \qquad (18)\end{aligned}$$

The first part of the expression is the sum of squares due to the causal effect of treatment (defined as treatment-induced variation, or *tiv*), and the second part is the sum of squares due to the mismatch error (defined as error-induced variation, or *eiv*). The third part of the expression has an expected value of zero (see Appendix B for details). Further define that:

$$TIV \equiv E\{tiv\} \qquad (19)$$

$$EIV \equiv E\{eiv\} \qquad (20)$$

Therefore, the expectation of the sum of squares of the predicted mediator can be expressed as TIV + EIV.

To see how error-induced variation causes instrumental variables estimates of β_{ca} to reflect the cross-sectional coefficient, β_{cs}, consider the following example. If β_{cs} is positive and the mean of M_* is higher for the treatment group than for the control group, the mean of Y_* is also likely to be higher for the treatment group. On the other hand, if the mean of M_* is lower for the treatment group, the mean value of Y_* is likely to be lower for the treatment group. In other words, the positive correlation between M_* and Y_* in cross-section produces a spurious positive correlation between *potential* values of $\frac{\Delta M}{\Delta T}$, the *estimated* impact of treatment on M, and $\frac{\Delta Y}{\Delta \mathrm{T}}$, the *estimated* impact of treatment on Y.[18] This spurious correlation induces a bias in their ratio, which is the Wald Estimator or its TSLS equivalent.

Figure 2 helps to illustrate this point graphically. Recall that the figure represents positive estimated treatment effects on M and Y, the ratio of which is an upward-sloping line. This slope is a Wald estimate of the causal effect of M on Y (or its two-stage least-squares equivalent). To the extent that the estimated treatment effects represent true treatment effects, their ratio represents the causal effect of M on Y or β_{ca}. This is how treatment-induced variation in the predicted mediator (*tiv*) comes into play. To the extent that the estimated treatment effects reflect mismatch error, their ratio represents the cross-sectional relationship between M_* and Y_* or β_{cs}. This is how error-induced variation (*eiv*) in the predicted mediator comes into play. Instrumental variables estimators for finite samples thereby reflect a mix of these factors.

To pursue this issue further, we express the treatment effects in Figure 2 as:

Treatment effect on the mediator

$$M_i \equiv M_{*i} + \pi T_i \qquad (21)$$

Treatment effect on the individual outcome

$$Y_i = Y_{*i} + \pi\beta_{ca}T_i \qquad (22)$$

Equation 21 states that the actual value of the mediator equals its counterfactual value plus the effect of treatment for treatment group members or plus zero for control group members. Equation 22 states that the actual value of the outcome equals its counterfactual value plus the effect of treatment for treatment group members or plus zero for control group members.

Substituting Equation 11 into Equation 22 yields:

$$Y_i = \alpha_{cs} + \beta_{cs}M_{*i} + \pi\beta_{ca}T_i + v_i \qquad (23)$$

Equation 23 indicates that the actual value of an individual outcome is a linear function of the counterfactual value of its mediator, times the cross-sectional coefficient, β_{cs}, plus the causal effect of treatment status on the outcome, $\pi\beta_{ca}$, for treatment group members or plus zero for control group

[18] The distribution of potential estimated values is the sampling distribution of their estimator.

members. Hence, systematic variation in the outcome reflects both the causal effect of treatment and the underlying cross-sectional coefficient.

Equation 23 implies that the difference in mean individual outcomes ($\bar{Y}_T - \bar{Y}_C$) for a treatment group and control group — the estimated effect of treatment on the outcome — is:

$$\bar{Y}_T - \bar{Y}_C = \frac{\Delta Y}{\Delta T} = \beta_{cs}(\bar{M}_{*T} - \bar{M}_{*C}) + \pi\beta_{ca} + (\bar{v}_T - \bar{v}_C) \qquad (24)$$

Equations 15 and 16 imply that the difference in mean mediator values ($\bar{M}_T - \bar{M}_C$) for a treatment group and control group — the estimated effect of treatment on the mediator — is:

$$\bar{M}_T - \bar{M}_C = \frac{\Delta M}{\Delta T} = (\bar{M}_{*T} - \bar{M}_{*C}) + \pi \qquad (25)$$

Hence, the Wald estimator (and its TSLS equivalent) of the effect of the mediator on the outcome is:

$$\hat{\beta}_{TSLS} = \frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{\beta_{cs}(\bar{M}_{*T} - \bar{M}_{*C}) + \pi\beta_{ca} + (\bar{v}_T - \bar{v}_C)}{(\bar{M}_{*T} - \bar{M}_{*C}) + \pi} \qquad (26)$$

Equation 26 illustrates that in finite samples (the only type to which researchers have access) the estimator is an amalgam of the true causal coefficient, β_{ca} , and the cross-sectional coefficient, β_{cs}.

Results in the Absence of Clustering

This section examines TSLS or Wald estimators from a randomized trial with a mediator, M, an individual outcome, Y, and a zero/one treatment-status indicator, T. N sample members are randomized in proportions $\bar{T}$ and $(1 - \bar{T})$ to treatment and control status, respectively.[19] These individuals are not clustered, and thus are statistically independent of each other. There is a true causal effect, π, of treatment on the mediator, a true causal effect, β_{ca} , of the mediator on the individual outcome, and a cross-sectional relationship, β_{cs} , between the mediator and outcome. In addition, there is a population variance, σ^2_{M*} , for counterfactual values of the mediator. The larger σ^2_{M*} is, the larger the treatment/control group mismatch is likely to be for a given sample and thus, the more error-induced variation there is likely to be in its predicted values of the mediator.

Finite Sample Bias in TSLS versus OLS Bias

Appendix B demonstrates that the median value of the sampling distribution of a TSLS estimator that is "just-identified" because its number of instruments equals its number of endogenous mediators is approximately:

[19] The symbol $\bar{T}$ is used to represent the proportion of sample members randomized to treatment because it equals the mean value of the zero/one treatment-status indicator for a study sample.

$$MEDIAN\{\hat{\beta}_{TSLS}\} \approx \frac{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}+[\sigma_{M*}^2]\beta_{cs}}{N\bar{T}(1-\bar{T})\pi^2+\sigma_{M*}^2} \quad (27)$$

This approximation is stated in terms of the *median* value of the sampling distribution of the estimator instead of its *mean or expected value* because its mean value does not exist (Basman, 1960, 1963; Bound, Jaeger, and Baker, 1995). Hence, for a just-identified TSLS model we discuss bias in terms of the median value of an estimator, or its "median bias".

Equation 27 indicates that as sample size goes to infinity (that is, as $N \to \infty$), MEDIAN$\{\hat{\beta}_{TSLS}\}$ converges to the true causal effect, β_{ca}, provided that $\pi \neq 0$. On the other hand, if $\pi = 0$, Equation 27 simplifies to

$$MEDIAN\{\hat{\beta}_{TSLS}\} \approx \frac{0+[\sigma_{M*}^2]\beta_{cs}}{0+\sigma_{M*}^2} = \beta_{cs} \quad (28)$$

Equation 28 demonstrates that if there is no treatment effect on the mediator, then any observed difference in the mean value of the mediator between the treatment and control group must necessarily be due to treatment-control mismatch of M. In this case, examining treatment-control differences in outcomes as a ratio to treatment-control differences in mediator levels is essentially equivalent to running a cross-sectional regression.

The preceding approximation indicates how treatment-induced variation (*tiv*) and error-induced variation (*eiv*) in predicted values of the mediator affect the resulting estimator. To see this, note the expected values of *tiv* and *eiv* are:

$$TIV = E\{tiv\} = N\bar{T}(1-\bar{T})\pi^2 \quad (29)$$

$$EIV = E\{eiv\} = \sigma_{M*}^2 \quad (30)$$

Equation 29 indicates that expected treatment-induced variation is proportional to sample size, the treatment/control group allocation, and the square of the true effect of treatment on the mediator. Consequently, other things being equal, one can increase this desirable variation by increasing sample size, increasing the causal effect of treatment on the mediator, and using a balanced sample (because $T = 0.5$, maximizes $\bar{T}(1-\bar{T})$). Equation 30 indicates that error-induced variation is proportional to the population variance of counterfactual values of the mediator. Usually, this undesirable variation is taken as given, although in principle, it could be reduced by conducting a trial within a particularly homogenous population.

Substituting Equations 29 and 30 into Equation 27 and rearranging terms yields the following expression (which is the same as Equation 10):

$$MEDIAN(\hat{\beta}_{TSLS}) \approx \left[\frac{TIV}{TIV+EIV}\right]\beta_{ca} + [\frac{EIV}{TIV+EIV}]\beta_{cs} \quad (31)$$

This result indicates that the median value of a TSLS or Wald estimator is approximately equal to an average of β_{ca} and β_{cs} that is weighted in proportion to TIV and EIV, respectively. For example, if TIV were only 20 percent of total variation in the predicted mediator, then only 20 percent of the weight of the average would be placed on the true causal effect, and 80 percent would be placed on the cross-sectional coefficient. To the extent that the cross-sectional coefficient is a biased estimate of the causal effect, there is bias in the TSLS or Wald estimator.

To see this, note that by definition:

$$BIAS_{TSLS} \equiv E\{\hat{\beta}_{TSLS}\} - \beta_{ca} \tag{32a}$$

or

$$MEDIANBIAS_{TSLS} \equiv MEDIAN\{\hat{\beta}_{TSLS}\} - \beta_{ca} \tag{32b}$$

Substituting Equation 31 into Equation 32b and simplifying terms yields:

$$MEDIANBIAS_{TSLS} \approx \left[\frac{EIV}{TIV+EIV}\right][\beta_{cs} - \beta_{ca}] \tag{33}$$

Thus, for example if EIV were 80 percent of total variation in the predicted mediator, the median bias of TSLS would equal 80 percent of the difference between β_{cs} and β_{ca}.

In the literature, finite sample bias is usually expressed in terms of "OLS bias," where:

$$BIAS_{OLS} \equiv E\{\hat{\beta}_{OLS}\} - \beta_{ca} \tag{34}$$

$$E\{\hat{\beta}_{OLS}\} = \beta_{cs} \tag{35}$$

and therefore:

$$BIAS_{OLS} = \beta_{cs} - \beta_{ca} \tag{36}$$

Substituting Equation 36 into Equation 33 yields the following relationship between finite sample bias for TSLS and OLS bias for one instrument and mediator:

$$MEDIANBIAS_{TSLS} \approx \left[\frac{EIV}{TIV+EIV}\right] BIAS_{OLS} \tag{37a}$$

Equation 37a — which is a very important result — indicates that finite sample bias for a TSLS or Wald estimator equals a fraction of the OLS bias that exists for a corresponding cross-sectional regression. This general point is widely noted in the literature.[20] That the particular fraction equals the

[20] The concepts of finite sample bias and weak instruments are so intertwined in the literature that it is difficult to determine which is most central.

error-induced proportion of total variation in predicted values of the mediator is not so well known. In the present hypothetical example — with 80 percent of the variation in predicted values of the mediator being error-induced — finite sample bias equals 80 percent of OLS bias.

One final point worth noting is that because OLS estimates are typically normally distributed, median OLS bias equals mean OLS bias and therefore:

$$MEDIANBIAS_{TSLS} \approx \left[\frac{EIV}{TIV+EIV}\right] MEDIANBIAS_{OLS} \qquad (37b)$$

Instrument Strength and Finite Sample Bias[21]

The existing literature focuses on the fact that finite sample bias is caused by "weak instruments."[22] A weak instrument is one that is weakly correlated with the mediator it is used to predict. The weaker the correlation is, the less predictive power the instrument has, and the weaker the instrument is. In our case, treatment status is the instrument. The greater the effect of treatment status on the mediator is, the stronger the correlation between treatment status and the mediator is, and the stronger treatment status is as an instrument.

It is often recommended that the strength of an instrument for a particular mediator be measured by the *population* F-value, F_{pop}, for the corresponding first-stage regression.[23] With a single instrument and mediator, the first-stage regression, Equation 8, is repeated below for convenience.

$$M_i = \mu + \pi T_i + \varepsilon_i \qquad \text{(8 restated)}$$

The *sample* F-statistic for this regression is used to test the statistical significance of the estimated coefficient, $\hat{\pi}$. This statistic equals the ratio of two estimates of variation in M per degree of freedom. The numerator of the ratio equals the predicted variation in M per instrument. This equals the total variation predicted by T divided by one for a single instrument, because each instrument represents a single degree of freedom. The numerator is thus equivalent to the total variation in predicted values of the mediator, which as noted earlier, equals (*tiv* + *eiv*). The denominator of the sample F-statistic equals the estimated variance, σ^2_{M*}, of unpredicted values of the mediator. This is equivalent to the error-induced variance in predicted values of the mediator, *eiv,* for a given sample. The sample F-statistic for a single instrument, $F^{(1)}_{sample}$, therefore equals:

$$F^{(1)}_{sample} = \frac{tiv+eiv}{eiv} \qquad (38)$$

[21] Appendix B derives the expressions presented in this section.

[22] Much of the recent literature on this topic is motivated by a study by Angrist and Krueger (1991) and a subsequent study by Bound, Jaeger, and Baker (1995)

[23] Bound, Jaeger, and Baker (1995).

The population F-value for a single instrument, $F_{pop}^{(1)}$, is the expected value of the corresponding sample F-statistic, which is *approximately* equal to the ratio of expected values of the numerator and denominator in Equation 38, or:

$$F_{pop}^{(1)} = E\left\{F_{sample}^{(1)}\right\} \approx \frac{E\{tiv+eiv\}}{E\{eiv\}} = \frac{TIV+EIV}{EIV} \tag{39}$$

This approximation rests on the fact that sample-based estimates of a population variance are quite accurate (they have little sampling variability) if they are based on more than about 20 degrees of freedom.[24] The inverse of $F_{pop}^{(1)}$ is thus:

$$\frac{1}{F_{pop}^{(1)}} \approx \frac{EIV}{TIV+EIV} \tag{40}$$

Equations 39 and 40 illustrate why $F_{pop}^{(1)}$ is a useful measure of instrument strength, or conversely, why $\frac{1}{F_{pop}^{(1)}}$ is a useful measure of instrument weakness. A large value for $F_{pop}^{(1)}$ implies a small value for $\frac{1}{F_{pop}^{(1)}}$, which in turn, implies that a small proportion of the variation in predicted values of the mediator is error-induced. This result implies that a large proportion of variation in predicted values of the mediator is treatment-induced. Hence, a large value for $F_{pop}^{(1)}$ indicates a strong instrument and a small value of $F_{pop}^{(1)}$ implies a weak instrument.

For example, if $F_{pop}^{(1)}$ equals 10 then $\frac{1}{F_{pop}^{(1)}}$ equals 1/10. This implies that one-tenth of the variation in predicted values of a mediator is error-induced and nine-tenths is treatment-induced, which typically would be considered a strong instrument. In contrast, if $F_{pop}^{(1)}$ equals 2 then $\frac{1}{F_{pop}^{(1)}}$ equals ½, which implies that half of the variation in predicted values of the mediator is error-induced and thus only half is treatment-induced. This typically would be considered a weak instrument.

When its implications for bias are considered, it becomes even clearer how $F_{pop}^{(1)}$ provides a useful measure of instrument strength. To see this, note that Equations 37 and 40 imply that:

[24] This point can be illustrated by the relationship that exists between a t distribution and a normal or z distribution. A t-statistic is the ratio of a sample-based parameter estimate to the sample-based estimate of its standard deviation (the square root of its variance). A z-statistic has the same numerator but assumes that the standard deviation (and thus variance) of the parameter is known. When the standard deviation of the estimator is estimated with very few degrees of freedom, the critical value for a t distribution (say for a two-tail hypothesis test at the 0.05 level of statistical significance) is much larger than that for a z distribution. This reflects the uncertainty — and thus variability — that exists for a sample-based estimate of a standard deviation or variance given very few degrees of freedom. For example, with only four degrees of freedom, the 0.05 two-tail critical value is 2.78 for a t-statistic versus 1.96 for a z-statistic. As the number of degrees of freedom (and thus sample size) increases, the critical value of a t-statistic rapidly approaches that of a z-statistic. For example, with 20 degrees of freedom the 0.05 two-tail critical value of a t-statistic is 2.09.

$$MEDIANBIAS_{TSLS} \approx \frac{1}{F_{pop}^{(1)}} BIAS_{OLS} \tag{41}$$

Equation 41 indicates that the median bias of a TSLS estimator with a single instrument and mediator is inversely proportional to the strength of the instrument used. Hence this bias is directly proportional to the weakness of the instrument.

For example, an $F_{pop}^{(1)}$ of 10 implies that finite sample bias equals one-tenth the magnitude of OLS bias, whereas an $F_{pop}^{(1)}$of 2 implies that finite sample bias equals half the magnitude of OLS bias. For this reason it is frequently recommended that a sample-based estimate of the F-value for the relevant first-stage regression be used to assess instrument's strength.

As a further point of reference, note that the preceding findings imply that:

$$MEDIAN\{\hat{\beta}_{TSLS}\} \approx \left[\frac{F_{pop}^{(1)}-1}{F_{pop}^{(1)}}\right] \beta_{ca} + \left[\frac{1}{F_{pop}^{(1)}}\right] \beta_{cs}. \tag{42}$$

Hence, the value of $F_{pop}^{(1)}$ determines the weights for β_{ca} and β_{cs} in the median value of a TSLS estimator. For example, if $F_{pop}^{(1)}$ equals 10, then the expected value of the estimator equals a weighted average of the true causal coefficient and the underlying cross-sectional coefficient, with a weight of 9/10 for the former and 1/10 for the latter. If $F_{pop}^{(1)}$ equals 2, both coefficients have a weight of ½. Consequently, the stronger an instrument is, the more weight it gives to the true causal coefficient.

Instrument Strength and Finite Sample Bias in Practice

At this point, the next natural question to ask is: How large must the effect of treatment on the mediator (π) and/or sample size (N) be in order for an instrument to be strong enough to reduce finite sample bias to an acceptable level? Unfortunately, a bias that is acceptable for one situation may be unacceptable for another. Hence, "acceptability" is not a universal parameter. Nevertheless, it is possible to illustrate in a simple and general way how the population F-value (F_{pop}) for an instrument varies as a function of sample size and the effect of treatment on the mediator expressed as a standardized mean difference *effect size.*

A standardized mean difference effect size or "effect size" for short, is a metric that is used widely in education research and related fields. It is defined as a treatment effect in its natural units divided by the standard deviation of the variable affected. The standard deviation used for this purpose is typically that for the counterfactual distribution of the affected variable, which in our case is the standard deviation of counterfactual values of the mediator or σ_{M*}. Consequently, the standardized mean difference effect size for a treatment effect on a mediator (ES_M) equals $\frac{\pi}{\sigma_{M*}}$.

Table 2

$\boldsymbol{F_{pop}}$ as a Function of Sample Size and Treatment Effect Size on Mediator: For a Balanced Sample Allocation ($\boldsymbol{\bar{T} = 0.5}$)

Treatment Effect Size on Mediator ($\boldsymbol{\pi/\sigma_{M*}}$)	**Sample Size (N)**				
	50	100	250	500	1000
0.2	1.5	2.0	3.5	6.0	11.0
0.4	3.0	5.0	11.0	21.0	41.0
0.6	5.5	10.0	23.5	46.0	91.0
0.8	9.0	17.0	41.0	81.0	161.0
1.0	13.5	26.0	63.5	126.0	251.0

Note that:

$$F_{pop} \approx \frac{TIV+EIV}{EIV} = 1 + \frac{TIV}{EIV}$$

$$\approx 1 + \frac{N\bar{T}(1-\bar{T})\pi^2}{\sigma_{M*}^2} = 1 + N\bar{T}(1-\bar{T})(\frac{\pi}{\sigma_{M*}})^2$$

$$\approx 1 + N\bar{T}(1-\bar{T})(ES_M)^2 \qquad (43)$$

Equation 43 indicates how instrument strength (F_{pop}) depends on sample size (N), the treatment/control group allocation ($\bar{T}(1-\bar{T})$), and the square of the effect size of treatment on the mediator (ES_M^2). Table 2 uses Equation 43 to illustrate this relationship for a balanced sample allocation (with $\bar{T} = 0.5$).

The existing literature (for example, see Stock and Yogo, 2005) often recommends that an instrument (or set of instruments) have an $F_{pop}$of at least 10 if it is to be used for a two-stage least-squares analysis. An F_{pop} of at least 10 implies that finite sample bias is no more than 1/10 of the bias from an ordinary least-squares cross-sectional regression analysis. Table 2 indicates that a sample of almost 1,000 units is required to meet this criterion if the treatment effect size on the mediator is 0.2 standard deviations. The required sample size declines dramatically as the treatment effect size increases so that fewer than 50 sample members are necessary when the treatment effect size is a full standard deviation.

Consider the implications of these findings for the two types of mediational analyses introduced earlier: individual-level analyses and setting-level analyses. An individual-level mediational analysis is one that examines the relationship between a mediator and an outcome that both vary across individuals.

Such an analysis might examine the relationship between students' academic engagement (the mediator) and their academic achievement (the outcome). Because educational research samples often contain hundreds or thousands of students, the treatment effect size for an individual mediator does not have to exceed 0.2 standard deviation in order for a proposed instrument (treatment status) to be strong enough for practical use (that is, for $F_{pop} \geq 10$).

Implications for setting-level mediational analyses are quite different, however, because research samples usually contain many fewer settings than individuals. Consider the archtypical setting-level mediational analysis in education research, with a measure of classroom instruction as the mediator and mean student achievement for each classroom as the outcome. With approximately 25 students per classroom, a research sample containing 2,500 students contains only 100 classrooms. This sample of classrooms places us in the portion of Table 2 that requires treatment effect sizes on the mediator of at least 0.6 standard deviation in order for F_{pop} to reach its desired value of 10 or greater. Hence, opportunities for conducting setting-level mediational analyses using instrumental variables may be more limited than those for conducting individual-level analyses. At the very least, they will require larger samples of settings.

To use the information provided by Equation 43 and Table 2 to assess the feasibility of instrumental variables analysis for a specific study or group of studies requires empirical knowledge about how large treatment effect sizes on mediators of interest are likely to be — which is beyond the scope of the present paper. However, we *can* offer an important point to consider when addressing this issue. This point derives from the fact that mediators are causally closer (more proximal) to treatments than are outcomes. Hence, the treatment effect size for a mediator will be larger than that for its related outcome. For example, the treatment effect size for a mediator will be twice that for its outcome if the correlation between them is 0.5. Consequently, effects sizes for mediators will be larger (sometimes by a lot) than those for outcomes, and researchers should set their expectations accordingly.

Results in the Presence of Clustering

The present study was motivated by our desire to use instrumental variables analysis with data for classrooms that are clustered in schools. For this analysis we needed to know how clustering affects instrument strength and finite sample bias, which to our knowledge, is not discussed in the literature. Consequently, we set out to explore the issue. Appendix C presents our results, which are summarized below.

Situation

Our results apply to situations where individuals or settings (units) are randomized and/or treated in interdependent groups (clusters). There are J clusters with a constant number of n units per cluster, and clusters are randomized in proportions $\bar{T}$ and $(1 - \bar{T})$ to a treatment group or a control group. This situation is more general than it seems because it approximates samples where the number of units per cluster varies and is represented by its harmonic mean.

Intuition

It is now well known that clustering does not affect the expected value of an estimated intervention effect from a randomized trial.[25] Hence, a trial that randomizes clusters provides unbiased estimates of treatment effects. This includes the first-stage regression in our TSLS analysis. It is also now well known that, other things being equal, clustering increases the variance of estimated intervention effects from a randomized trial. Hence, clustering increases the variance of estimates of our first-stage regression coefficient.

Consider a sample of 1,000 students (units) from 10 schools (clusters), with 100 students per school. Regardless of whether students or schools are randomized, the expected value of the estimated intervention effect equals the true intervention effect. This is because randomization ensures that each student (or school) has the same probability of being assigned to treatment. If randomization were repeated an infinite number of times, the mean counterfactual value of the dependent variable (M_* in our first-stage regression) would be the same for the treatment group and control group. In other words, the expected value of M_* is the same for the treatment group and control group, regardless of whether schools or students are randomized.

However, a treatment group and control group are likely to be mismatched on M_* for a given randomization (draw). For example, if each of our 1,000 students were randomized independently, the potential for a large treatment/control group mismatch is much less than if the 10 schools they attended were randomized. Hence, for a given total number of units, the variance of potential mismatches is larger (often by a lot) when clusters instead of units are randomized.

As noted earlier, the variance of the potential mismatch *is* the variance of the estimated intervention effect on the mediator. Hence, for a given total number of units, the variance of an estimated intervention effect on a mediator is larger with clustering than without it, which increases likely mismatch error for the mediator. This means that, other things being equal, clustering increases error-induced variation in predicted values of a mediator, or:

$$EIV_{CL} \geq EIV \tag{44}$$

where EIV_{CL} is the expected value of error-induced variation in predicted values of the mediator with clustering, and EIV is its counterpart without clustering. In contrast, because clustering does not affect the expected value of an estimated first-stage intervention effect:

$$TIV_{CL} = TIV \tag{45}$$

[25] The statistical properties of cluster-randomized designs have become widely recognized only recently. See Bloom (2005) for discussion.

Consequently, clustering increases error-induced variation relative to treatment-induced variation, thereby weakening the predictive power of a first-stage instrument and increasing bias from TSLS or Wald estimation.

Model

To represent clustering in our TSLS analysis, the first-stage regression becomes:

$$M_{ij} = \mu + \pi T_j + e_j + \varepsilon_{ij} \qquad (46)$$

where:

$$E\{\hat{\pi}_{CL}\} = \pi \qquad (47)$$

$$VAR(e_j) \equiv \tau_{M*}^2 \qquad (48)$$

$$VAR(\varepsilon_{ij}) \equiv \theta_{M*}^2 \qquad (49)$$

Subscripts i and j denote the i^th unit in the j^th cluster; e_j is an independent random error for the j^th cluster; and ε_{ij} is an independent random error for the i^th unit in the j^th cluster.

It is common practice to report the relationship between unit-level and cluster-level variance components (*aka* within and between cluster variation) as an intra-class correlation, ρ. This parameter is defined as the ratio of the cluster-level variance component to the sum of the cluster-level and unit-level variance components; and this sum equals the *total* variance of units within and between clusters. In the present case:

$$\rho_{M*} = \frac{\tau_{M*}^2}{\tau_{M*}^2 + \theta_{M*}^2} \qquad (50)$$

The intra-class correlation provides a measure of the degree to which units are clustered. The larger the value of this parameter is, the more clustered (or segregated) units are. If the intra-class correlation equals zero, units are not at all clustered and none of the unit variation is between clusters. If the intra-class correlation equals one, units are fully clustered and all of the unit variation is between clusters.

Consider how the statistical properties of our first-stage regression in the presence of clustering (Equation 46) differ from those in the absence of clustering (Equation 8). For this comparison, it is necessary to hold constant the total variance of the counterfactual values of the mediator. This implies that the total unit variance without clustering, σ_{M*}^2, equals the total unit variance with clustering, $(\tau_{M*}^2 + \theta_{M*}^2)$. The only difference between these two situations is their distribution of the unit variation between and within clusters. Consequently:

$$\rho_{M*} = \frac{\tau_{M*}^2}{\sigma_{M*}^2} \qquad (51)$$

and

$$1 - \rho_{M*} = \frac{\theta^2_{M*}}{\sigma^2_{M*}} \tag{52}$$

Appendixes B and C demonstrate that the variance of an estimated intervention effect on a mediator without clustering, $VAR(\hat{\pi})$, and its counterpart with clustering, $VAR(\hat{\pi}_{CL})$ are:

$$VAR(\hat{\pi}) = [\frac{\boldsymbol{\sigma^2_{M*}}}{\boldsymbol{\bar{T}(1-\bar{T})}}][\frac{1}{N}] \tag{53}$$

and

$$VAR(\hat{\pi}_{CL}) = [\frac{1}{\bar{T}(1-\bar{T})}][\frac{\tau^2_{M*}}{J} + \frac{\theta^2_{M*}}{nJ}\}$$

$$= [\frac{\sigma^2_{M*}}{\bar{T}(1-\bar{T})}][\frac{\rho_{M*}}{J} + \frac{1-\rho_{M*}}{N}]$$

$$= [\frac{\boldsymbol{\sigma^2_{M*}}}{\boldsymbol{\bar{T}(1-\bar{T})}}][\rho_{M*}\left(\frac{1}{J}\right) + (1-\rho_{M*})(\frac{1}{N})] \tag{54}$$

The bolded terms in Equations 53 and 54 are the same. However in Equation 53 these terms are multiplied by the inverse of the total number of sample units (N), whereas in Equation 54 they are multiplied by a weighted average of the inverse of the number of sample clusters (1/J) and the inverse of the number of sample units (1/N). Because there are fewer (often many fewer) clusters than units, the product in Equation 54 is larger than that in Equation 53 if the intra-class correlation is not zero. Hence, the variance of the estimator for a first-stage regression coefficient is larger with clustering than without it. Furthermore, as clustering (ρ_{M*}) increases, the weight placed on the inverse of J increases, and the product of Equation 54 increases. In other words, as clustering increases, the variance of the first-stage estimator increases.

Results

Appendix C demonstrates that in the presence of clustering:

$$MEDIAN\{\hat{\beta}_{TSLS(CL)}\} \approx [\frac{TIV}{TIV+EIV_{CL}}]\beta_{ca} + [\frac{EIV_{CL}}{TIV+EIV_{CL}}]\beta_{cs} \tag{55}$$

$$MEDIANBIAS_{TSLS(CL)} \approx \left[\frac{EIV_{CL}}{TIV+EIV_{CL}}\right](\beta_{cs} - \beta_{ca}) \tag{56}$$

$$F^{(1)}_{pop(CL)} \approx \frac{TIV+EIV_{CL}}{EIV_{CL}} \tag{57}$$

Equations 55 to 57 differ from their counterparts without clustering only with respect to the fact that EIV_{CL}(with clustering) replaces EIV(without clustering) because clustering affects only error-induced variation in predicted values of the mediator.

To see how this influences finite sample bias and instrument strength, note that as demonstrated by Appendix C:

$$EIV_{CL} = \boldsymbol{\sigma^2_{M*}}[1 + (n - 1)\rho_{M*}] \tag{58}$$

and recall that:

$$EIV = \boldsymbol{\sigma^2_{M*}} \tag{59}$$

Hence, clustering increases the error-induced variation in predicted values of a mediator by a factor of $[1 + (n - 1)\rho_{M*}]$. This implies that error-induced variation increases with an increase in the intraclass correlation and with an increase in the number of units per cluster, for a given total number of units and counterfactual mediator variation.

Equation 58 and 59 in conjunction with Equation 55 imply that, other things being equal, clustering reduces the relative weight placed by TSLS estimators on the true causal effect (β_{ca}) of a mediator. It therefore increases the relative weight of the underlying cross-sectional coefficient (β_{cs}). Equations 58 and 59 in conjunction with Equation 56 imply that: *Other things being equal, clustering increases bias in TSLS or Wald estimators.* Equations 58 and 59 in conjunction with Equation 57 imply that, other things being equal, clustering reduces the F-value for an instrument in a first-stage regression.[26] Hence, clustering reduces the strength of an instrument.

The parallel effects of clustering on instrument strength and finite sample bias imply that the relationship between them is the same in the presence or absence of clustering. Consequently, the first-stage F-value has the same implications for finite sample bias in either case, so that:

$$\begin{aligned} MEDIANBIAS_{TSLS(CL)} &\approx \left[\frac{1}{F^{(1)}_{pop(CL)}}\right][\beta_{cs} - \beta_{ca}] \\ &\approx [\frac{1}{F^{(1)}_{pop(CL)}}]BIAS_{OLS} \end{aligned} \tag{60}$$

[26] Note that $F^{(1)}_{pop(CL)} = \frac{TIV+EIV_{CL}}{EIV_{CL}} = \frac{TIV}{EIV_{CL}} + \frac{EIV_{CL}}{EIV_{CL}} = \frac{TIV}{EIV_{CL}} + 1$. Therefore, as EIV_{CL} increases $\frac{TIV}{EIV_{CL}}$ and thus $F^{(1)}_{pop(CL)}$decrease.

Clustering, Instrument Strength, and Finite Sample Bias in Practice

Now consider how clustering affects the relationship between instrument strength, finite sample bias, sample size, and treatment effect size for a mediator. Note that:

$$F_{pop} = 1 + \frac{Jn\bar{T}(1-\bar{T})\pi^2}{[1+(n-1)\rho_{M*}]\sigma_{M*}^2}$$

$$= 1 + \frac{Jn\bar{T}(1-\bar{T})ES_M^2}{1+(n-1)\rho_{M*}} \qquad (61)$$

Equation 61 (with clustering) is the same as Equation 43 (without clustering) except for the "cluster design effect" $1+(n-1)\rho_{M*}$. Because this effect is positive whenever the intra-class correlation (ρ_{M*}) exceeds zero, it reduces F_{pop} for a given sample size (Jn), sample allocation ($\bar{T}(1-\bar{T})$, and treatment effect size (ES_{M*}). To explore the implications of Equation 61, consider the following scenario.

We are studying the effects of an educational intervention on third-grade reading achievement by randomizing half of the elementary schools in our sample to a treatment group that receives the intervention and the other half to a control group that does not receive the intervention. The intervention is a form of professional development that is designed to improve teachers' reading instruction in ways that are intended to increase student engagement, which in turn, is expected to increase student reading achievement.

We will conduct two types of mediational analyses: one at the individual (student) level and one at the setting (classroom) level. The outcome of interest for both analyses will be student reading achievement, measured by scores on a standardized test. For the individual-level analysis, our mediator will be student engagement, measured by student survey responses. For our setting-level analysis, the mediator will be teacher reading instruction, measured by classroom observations. There are three third-grade classrooms per school and 25 students per classroom.

Table 3 presents values of F_{pop} for the individual-level mediational analysis (with 75 students per school) given the number of clusters (schools) in the study, the treatment effect size on the mediator, and the intra-class correlation. Table 4 presents corresponding information for the setting-level mediational analysis (with three classrooms per school). Comparisons of the two tables indicate that F_{pop} is likely to be much larger for the individual-level analysis than for the setting-level analysis. Hence, the individual-level analysis probably will be less susceptible to finite sample bias.

Findings in Table 3 for the individual-level analysis exhibit the expected pattern of increasing values of F_{pop} with increases in the number of clusters and treatment effect size. Furthermore, there is a marked decline in values for F_{pop} as clustering (measured by the intra-class correlation) increases.

Table 4, which has much smaller values for F_{pop}, also exhibits a consistent pattern of increases in these values with increases `in the number of clusters and treatment effect size. In addition, the table

Table 3

F_{pop} as a Function of the Number of Clusters and Treatment Effect Size on Mediator: For a Balanced Sample Allocation ($\bar{T} = 0.5$) With 75 Units per Cluster

Treatment Effect Size on Mediator (π/σ_{M*})	**Number of Clusters (J)**				
	20	40	60	80	100
			$\rho_{M*} = 0.05$		
0.2	4.2	7.4	10.6	13.8	17.0
0.4	13.8	26.5	39.3	52.1	64.8
0.6	29.7	58.5	87.2	115.9	144.6
0.8	52.1	103.1	154.2	206.3	256.3
1.0	80.8	160.6	240.4	320.2	399.9
			$\rho_{M*} = 0.15$		
0.2	2.2	3.5	4.7	6.0	7.2
0.4	6.0	10.9	15.9	20.8	25.8
0.6	12.2	23.3	34.5	45.6	56.8
0.8	20.8	40.7	60.5	80.3	100.2
1.0	32.0	63.0	94.0	125.0	156.0
			$\rho_{M*} = 0.25$		
0.2	1.8	2.5	3.3	4.1	4.9
0.4	4.1	7.2	10.2	13.3	16.4
0.6	7.9	14.9	21.8	28.7	35.6
0.8	13.3	25.6	37.9	50.2	62.5
1.0	20.2	39.5	58.7	77.9	97.2

exhibits a pattern of decreasing values for F_{pop} with increasing values of the intra-class correlation. However, the proportional effect of increased clustering on decreased values of $F_{pop}$is less pronounced for the setting-level analysis in Table 4 than for the individual-level analysis in Table 3. This is because there are far fewer units per cluster for the setting-level analysis (where n = 3) than for the individual-level analysis (where n = 75).

On balance then, the effect of clustering is likely to be greater for an individual-level mediational analysis than for a setting-level mediational analysis. However, a treatment-based instrument is likely to be much weaker for a setting-level analysis than for an individual-level analysis. Nevertheless, there seem to be realistic situations for both types of analyses in which instrument strength attains its prescribed minimum level (with $F_{pop} \geq 10$.)

Table 4

F_{pop} as a Function of the Number of Clusters and Treatment Effect Size on Mediator: For a Balanced Sample Allocation ($\bar{T} = 0.5$)
With 3 Units per Cluster

Treatment Effect Size on Mediator (π/σ_{M*})	**Number of Clusters (J)**				
	20	40	60	80	100
			$\rho_{M*} = 0.05$		
0.2	1.6	2.1	2.6	3.2	3.7
0.4	3.2	5.4	7.6	9.7	11.9
0.6	5.9	10.8	15.7	20.6	25.6
0.8	9.7	18.5	27.2	35.9	44.6
1.0	14.6	28.3	41.9	55.6	69.2
			$\rho_{M*} = 0.15$		
0.2	1.5	1.9	2.4	2.9	3.3
0.4	2.9	4.7	6.5	8.4	10.2
0.6	5.2	9.3	13.5	17.6	21.8
0.8	8.4	15.8	23.2	30.5	37.9
1.0	12.5	24.1	35.6	47.2	58.7
			$\rho_{M*} = 0.25$		
0.2	1.4	1.8	2.2	2.6	3.0
0.4	2.6	4.2	5.8	7.4	9.0
0.6	4.6	8.2	11.8	15.4	19.0
0.8	7.4	13.8	20.2	26.6	33.0
1.0	11.0	21.0	31.0	41.0	51.0

Part 3

Finite Sample Bias with a Single Mediator and Multiple Randomized Instruments

This section considers finite sample bias for a single mediator that is predicted by multiple instruments that are created from treatment indicators for multiple studies, sites, or subgroups. We refer interchangeably to these sample subdivisions as "*strata*."

Situation, Model, and Questions

Consider a randomized trial for each of K strata, with treatment status, T, mediator, M, and outcome, Y. Equations 62 and 63 below provide a conceptual model of the first and second stages of the corresponding TSLS analysis; Equation 64 represents predicted values of the mediator.

$$M_{ik} = \mu_k + \pi_k T_{ik} + \varepsilon_{ik} \tag{62}$$

$$Y_{ik} = \alpha_k + \beta T_{ik} + \nu_{ik} \tag{63}$$

$$\widehat{M}_{ik} = \hat{\mu}_k + \hat{\pi}_k T_{ik} \tag{64}$$

where the subscript, ik, represents the i^th unit in the k^th stratum. To simplify the discussion we first assume that units are not clustered and then add clustering later. To simplify further, we assume that all strata have the same total number of units, $(N_k = N/K)$, proportion of units randomized to treatment, $\bar{T}_k$, and variance of counterfactual mediator values, $\sigma^2_{M*(K)}$.

In the present case it is possible to create K instrumental variables by interacting treatment status, T, with a dichotomous indicator for each stratum and pooling data across strata. Equations 65 to 67 below represent this operational model. Note that they index parameters and strata indicators with a subscript or superscript, m, to distinguish them from the subscript, k, that identifies the stratum for each unit, i. This cumbersome distinction is not essential for understanding what follows.

$$M_{ik} = \sum_{m=1}^{K} \mu_m S_k^{(m)} + \sum_{m=1}^{K} \pi_m S_k^{(m)} T_{ik} + \varepsilon_{ik} \tag{65}$$

$$Y_{ik} = \sum_{m=1}^{K} \alpha_m S_k^{(m)} + \beta M_{ik} + \upsilon_{ik} \tag{66}$$

$$\widehat{M}_{ik} = \sum_{m=1}^{K} \hat{\mu}_m S_k^{(m)} + \sum_{m=1}^{K} \hat{\pi}_m S_k^{(m)} T_{ik} \tag{67}$$

where $S_{ik}^{(m)}$equals one when m equals k and zero otherwise. Note that in both Equation 65 and Equation 66, the stratum fixed effects are included in the model. This ensures that after pooling strata, the variance of the mediator in the control group (conditional on stratum dummies) will continue to be σ^2_{M*} (Strata fixed effects also would be included in analyses that use a single instrument and pool data across strata).

One reason for constructing multiple strata-specific instruments based on treatment status instead of using a single full-sample instrument is that the predictive power of multiple instruments can be greater than that for a single instrument, if the true impact of treatment on the mediator varies sufficiently across strata. In this case, using multiple instruments can reduce finite sample bias in a TSLS analysis. But if the variation across strata in the effect of treatment on the mediator is not sufficient, using multiple instruments can increase finite sample bias — often by a lot. The present section addresses this issue.

A second reason for using multiple instruments is to separate the effects of multiple mediators for an intervention and outcome. This type of analysis is much more complex than that for a single mediator and raises issues that are far beyond the scope of the present analysis. We shall address these issues in future research.

Specifically, the present section addresses the following questions:

- What happens when multiple instruments are used but the effect of treatment on the mediator is constant across strata?
- How much variation across strata in the treatment effect on the mediator is required for multiple mediators to reduce finite sample bias?
- How does clustering affect answers to the preceding questions?

Constant Treatment Effects and No Clustering

We begin with the properties of multiple strata-specific instruments for a single mediator, given a constant treatment effect on the mediator and no clustering. Appendix D derives the findings presented.

Treatment-Induced Variation versus Error-Induced Variation

It is perhaps easiest to understand the K-strata situation by considering each stratum as a separate randomized trial and pooling findings across them. For a given stratum, N/K units are randomized in proportions $\bar{T}$ and $(1-\bar{T})$ to treatment and control status, and the variance of counterfactual values of the mediator is σ^2_{M*}. Hence, the only way that stratum k differs from the full sample is that it has (1/K)th of the sample members (units).

In terms of treatment-induced and error-induced variation of predicted mediator values for the kth stratum, $tiv^{(k)}$ and $eiv^{(k)}$, where k = 1, 2,…,K, this implies that:

$$E\{tiv^{(k)}\} = TIV^{(k)} = \left[\frac{N}{K}\right]\bar{T}(1-\bar{T})\pi^2 \qquad (68)$$

and

$$E\{eiv^{(k)}\} = EIV^{(k)} = \sigma^2_{M*} \qquad (69)$$

Summing these expected values over the K strata is equivalent to multiplying them by K. Hence, the full-sample expected values of treatment-induced and error-induced variation are:

$$TIV^{(K)} = E\left\{\sum_{k=1}^{K} tiv^{(k)}\right\} = K \cdot TIV^{(k)} = N\bar{T}(1-\bar{T})\pi^2 \qquad (70)$$

$$EIV^{(K)} = E\left\{\sum_{k=1}^{K} eiv^{(k)}\right\} = K \cdot EIV^{(k)} = K \cdot \sigma_{M*}^2 \qquad (71)$$

Hence, $TIV^{(K)}$ for K strata-specific, treatment-based instruments equals $TIV^{(1)}$ for a single full-sample, treatment-based instrument, or:

$$TIV^{(K)} = TIV^{(1)} \qquad (72)$$

However, because the sample for each stratum is 1/K$^{\text{th}}$ the size of the full sample, $EIV^{(K)}$ for K instruments is K times $EIV^{(1)}$ for a single instrument, or:

$$EIV^{(K)} = K \cdot EIV^{(1)} \qquad (73)$$

This result occurs because the estimated effect of treatment on the mediator for each stratum, $\hat{\pi}_k$, is based only on 1/K$^{\text{th}}$ of the full sample. Hence, the variance of its estimator — which is the variance of potential mismatch error — is K times the variance of the corresponding full-sample estimator for a single instrument. Equations 72 and 73 are the key to understanding how the properties of multiple instruments affect their overall strength and thus how they affect finite sample bias.

Instrument Strength and Finite Sample Bias

Appendix D demonstrates that for our K strata-specific instruments:

$$F_{pop}^{(K)} = \frac{TIV^{(K)} + EIV^{(K)}}{EIV^{(K)}} \qquad (74)$$

and for a single full-sample instrument:

$$F_{pop}^{(1)} = \frac{TIV^{(1)} + EIV^{(1)}}{EIV^{(1)}} \qquad (75)$$

Substituting Equations 72 and 73 into Equation 74 and rearranging terms yields:

$$F_{pop}^{(K)} = \frac{TIV^{(1)} + K \cdot EIV^{(1)}}{K \cdot EIV^{(1)}}$$

$$= \frac{TIV^{(1)}}{K \cdot EIV^{(1)}} + 1 \qquad (76)$$

Rearranging terms in Equation 75 yields:

$$F_{pop}^{(1)} = \frac{TIV^{(1)}}{EIV^{(1)}} + 1 \qquad (77)$$

Therefore:

$$F_{pop}^{(K)} < F_{pop}^{(1)} \tag{78}$$

In words, the overall "strength" of the set of K strata-specific instruments is unambiguously less than that of a single full-sample instrument when the effect of treatment on the mediator is constant across strata.

To ascertain how this difference in the strength of a set of K instruments versus that of a single instrument affects their relative bias for a TSLS estimator, it is first necessary to put them on a common basis of comparison. As noted earlier, the expected value or mean of the sampling distribution for a single-instrument TSLS estimator (which is just identified) does not exist. Thus we must assess its median bias, which can be approximated by Equation 41 restated below.

$$MEDIANBIAS_{TSLS}^{(1)} \approx \frac{1}{F_{pop}^{(1)}} BIAS_{OLS} \tag{41 restated}$$

The expected value or mean of a TSLS estimator with K instruments and a single endogenous mediator (which is overidentified) does exist (for example, see Bound et al., 1995). Hence it is possible to assess the mean bias of this estimator, which can be approximated as follows:

$$BIAS_{TSLS}^{(K)} \approx \frac{1}{F_{TSLS}^{(K)}} BIAS_{OLS} \tag{79}$$

It is also the case that an overidentified TSLS estimator with K instruments has an asymptotically normal sampling distribution (Angrist and Pischke, 2009, p. 140). Thus its asymptotic mean equals its asymptotic median and:

$$MEDIANBIAS_{TSLS}^{(K)} \approx \frac{1}{F_{TSLS}^{(K)}} BIAS_{OLS} \tag{80}$$

Because our set of K instruments are weaker than a single instrument in the present example ($F_{pop}^{(K)} < F_{pop}^{(1)}$) it follows that:

$$MEDIANBIAS_{TSLS}^{(K)} > MEDIANBIAS_{TSLS}^{(1)} \tag{81}$$

Varying Treatment Effects and No Clustering

Now consider what happens when the impact of treatment on a mediator, π_k , varies across strata. To simplify the discussion, we focus on the pattern of treatment-effects which has a constant difference, ϕ, between adjacent treatment effects when they are rank-ordered from least positive to most positive. In symbols:

$$\pi_k - \pi_{k-1} = \phi \tag{82}$$

Appendix D demonstrates that the mean and variance of treatment effects for this distribution are:

$$E\{\pi_k\} = \pi_1 + [\tfrac{K-1}{2}]\phi \tag{83}$$

and

$$VAR(\pi_k) = [\tfrac{K^2-1}{12}]\phi^2 \tag{84}$$

The key to understanding how this distribution of treatment effects on the mediator influences finite sample bias for a TSLS analysis based on K strata-specific instruments is to understand how they affect treatment-induced variation and error-induced variation in predicted values of the mediator. For K instruments, Appendix D demonstrates that:

$$E\{tiv^{(K)}\} \equiv TIV^{(K)} = N\bar{T}(1-\bar{T})[\tfrac{1}{K}]\textstyle\sum_{k=1}^{K}\pi_k^2$$

$$= N\bar{T}(1-\bar{T})[\pi_1^2 + (K-1)\pi_1\phi + \phi^2\tfrac{(K-1)(2K-1)}{6}] \tag{85}$$

and

$$E\{eiv^{(K)}\} \equiv EIV^{(K)} = K \cdot \sigma_{M*}^2 \tag{86}$$

For a single full-sample instrument, Appendix D demonstrates that:

$$E\{tiv^{(1)}\} = TIV^{(1)} = N\bar{T}(1-\bar{T})(\pi_1 + \tfrac{K-1}{2}\phi)^2 \tag{87}$$

$$E\{eiv^{(1)}\} = EIV^{(1)} = \sigma_{M*}^2 \tag{88}$$

Equations 86 and 88 indicate that K instruments produce K times as much error-induced variation in predicted values of the mediator as does a single instrument. This is because the estimated treatment effect for each of the K instruments is based on 1/K$^{\text{th}}$ of the sample, whereas that for a single instrument is based on the full sample. In symbols:

$$EIV^{(K)} = K \cdot EIV^{(1)} \tag{89}$$

This result applies whether or not the treatment effect on the mediator varies across strata.

In order for the *strength* of the set of K instruments to exceed that of a single instrument — and therefore reduce finite sample bias — the treatment-induced variation from K instruments (Equation 85) must exceed that for a single instrument (Equation 87) by an amount that more than offsets the increased error-induced variation produced by K instruments instead of one. In symbols:

$$TIV^{(K)} - TIV^{(1)} > EIV^{(K)} - EIV^{(1)}$$

$$> K \cdot \sigma_{M*}^2 - \sigma_{M*}^2$$

$$> (K-1)\sigma_{M*}^2 \qquad (90)$$

If this condition is met, the first-stage F-value for K instruments will be larger than that for a single instrument, and finite sample bias for K instruments will be smaller than for a single instrument. If this condition is not met, the first-stage F-value for K instruments will be smaller than that for a single instrument, and finite sample bias for K instruments will be larger than that for a single instrument. If the two sides of Equation 90 are equal, then instrument strength and finite sample bias will be the same for K instruments or a single instrument. Appendix D demonstrates that the condition in Equation 90 implies that:

$$VAR(\pi_k) > (K-1)(E\{\pi_k\})^2 \qquad (91)$$

Equation 91 states that in order to justify using K instruments instead of one for a single mediator, the variance across strata of treatment effects on the mediator must exceed $(K-1)$ times the square of the mean effect of treatment on the mediator.

In theory, the preceding condition is met when the population F-value for K instruments exceeds that for a single instrument (see Appendix D). Thus in practice, researchers can use the sample F-statistic as a guide for assessing whether or not the condition is likely to be met. This practice is widely recommended in the literature (Bound, Jaeger, and Baker, 1995).

In thinking about when the condition in Equation 91 is likely to be met, it is useful to simplify the situation even further by considering the case of two equal-size strata. Substituting Equations 83 and 84 into Equation 91, setting K equal to 2, and rearranging terms yields:

$$VAR(\pi_k) > (2-1)[E\{\pi_k\}]^2$$

$$\phi^2[\tfrac{4-1}{12}] > [\pi_1 + \tfrac{1}{2}\phi]^2$$

$$\tfrac{1}{4}\phi^2 > \pi_1^2 + \phi\pi_1 + \tfrac{1}{4}\phi^2$$

$$0 > \pi_1^2 + \phi\pi_1 \quad -\phi\pi_1 > \pi_1^2 \qquad (92)$$

Equation 92 implies that in order for the condition in Equation 91 to be met for two equal-size strata, the effect of treatment on the mediator must be positive for one stratum and negative for the other. This makes it possible for the variance of the treatment effects to be large enough relative to their mean to outweigh the increase in error-induced variation produced by estimating separate treatment effects for each half of the sample.

The Effect of Clustering

As it was for a single instrument and mediator, the effect of clustering for multiple instruments and a single mediator is to increase the variance of estimates of the impact of treatment on the mediator.

Random mismatch error is increased accordingly, which in turn, increases error-induced variation in predicted mediator values. This increase is the same whether the treatment effect on the mediator is constant or varies across clusters. Other things being equal, clustering inflates error-induced variation as follows:

$$EIV_{CL} = EIV[1 + (n - 1)\rho_{M*}] \qquad (93)$$

On the other hand, clustering does not affect treatment-induced variation in predicted values of a mediator so that TIV_{CL} equals TIV. Consequently, other things being equal, clustering reduces the overall strength of a set of K instruments and thereby increases finite sample bias. However, clustering does not affect the trade-off between using multiple instruments or a single instrument for a single mediator.

Part 4

Next Steps

This paper represents the first step in our exploration of instrumental variables analysis with randomized trials to study causal relationships between features of settings (mediators) and outcomes for individuals. The paper: (1) highlights the main problems with commonly used alternatives for conducting such causal analyses, (2) introduces the instrumental variables approach, (3) examines the problem of finite sample bias for such analyses given a single mediator and instrument, (4) examines this problem for a single mediator and multiple instruments, and (5) compares finite sample bias of the two-stage least-squares estimator in clustered and unclustered designs.

However, more and more complex issues must be considered in order to complete a comprehensive assessment of the instrumental variables approach being considered. And it is these issues to which we will turn next. Among them are:

1. How does instrument strength affect estimated standard errors for single instruments and multiple instruments?
2. How does failure of the exclusion principle affect results, and how sensitive are results likely to be to failures of assumption that are of a magnitude that might be expected in practice?
3. Under what conditions can multiple instruments be used to estimate the separate causal effects of multiple mediators, and when are these conditions likely to be met in practice?
4. What are the properties of other estimation procedures for instrumental variables analyses (for example, limited information maximum likelihood [Anderson and Rubin, 1949] and random effects maximum likelihood [Chamberlain and Imbens, 2004]) for the types of analyses being considered?

Appendix A

Attenuation Bias in OLS Estimators

This appendix demonstrates the existence and examines the magnitude of attenuation bias in ordinary least squares (OLS) estimators of the relationship between a mediator and outcome. This bias is caused by random measurement error in the mediator. To simplify the discussion, it assumes that there is no omitted variable bias or simultaneity bias. Proofs are presented for the following propositions:

Proposition A.1: Consider the following regression model:

$$Y_i = \alpha + \beta_{ca} M_{true,i} + \nu_i \qquad \text{(A.1)}$$

where:

$Y_i \; and \; M_{true,i}$ = true values of the outcome and mediator for individual i,

β_{ca} = the causal regression coefficient for the mediator, and

ν_i = a random error that is independently and identically distributed.

Using OLS to estimate this regression, absent omitted variables, simultaneity, or covariates, and assuming purely random measurement error for the mediator, the probability limit of the OLS estimator for the causal relationship between the outcome, Y_i , and the mediator, M_i, is:

$$plim(\hat{\beta}_{OLS}) = \lambda\beta_{ca} \qquad \text{(A.2)}$$

where λ is the reliability of the observed mediator and is defined as:

$$\lambda \equiv \frac{VAR(M_{true,i})}{VAR(M_{true,i})+VAR(M_{error,i})} \qquad \text{(A.3)}$$

Proposition A.2: Consider the following augmented regression:

$$Y_i = \alpha + \beta_{ca} M_{true,i} + \beta_1 X_{1,i} + \nu_i \qquad \text{(A.4)}$$

where $X_{1,i}$ is a correctly measured covariate for individual I; β_{ca} is the causal relationship between the outcome and the mediator, conditioning on $X_{1,i}$; and $\beta_1$is the relationship between the outcome and the covariate $X_{1,i}$. All other variables are the same as defined in Equation A.1.

Absent omitted variables bias or simultaneity bias, assuming purely random measurement error for the mediator, and including one correctly measured covariate (X_1) in the regression model, (where X_1 is uncorrelated with the measurement error for the mediator and the residuals in the regression), the probability limit of the OLS estimator is:

$$plim(\hat{\beta}_{OLS}) = \left[\frac{\lambda - R^2_{MX1}}{1 - R^2_{MX1}}\right]\beta_{ca} \qquad \text{(A.5)}$$

where λ is defined as in Equation A.3 and R^2_{MX1} is the square of the correlation (R-squared) between the observed mediator and the covariate, X_1.

Proofs of Propositions A.1 and A.2

Proof of Proposition A.1

Consider the following regression:

$$Y_i = \alpha + \beta_{ca} M_{true,i} + \nu_i \qquad \text{(A-A.1)}$$

where:

Y_i and $M_{true,i}$ = true values of the outcome and mediator for individual i,

β_{ca} = the causal regression coefficient for the mediator, and

ν_i = a random error term that is independently and identically distributed.

Note that the coefficient for the mediator has causal interpretation because in this appendix we assume no omitted variables bias or simultaneity bias.

Now suppose that $M_{true,i}$ is measured imprecisely by M_i, such that:

$$M_i = M_{true,i} + M_{error,i} \qquad \text{(A-A.2)}$$

Where the measurement error for individual i, $M_{error,i}$, is purely random with mean zero and variance $VAR(M_{error,i})$, and it is uncorrelated with $M_{true,i}$ and the regression error ν_i.

Because M_i, not $M_{true,i}$, is observed, the regression equation is based on M_i. Substituting Equation A-A.2 into equation A-A.1 yields:

$$\begin{aligned} Y_i &= \alpha + \beta_{ca} M_i - \beta_{ca} M_{error,i} + \nu_i \\ &= \alpha + \beta_{ca} M_i + \omega_i \qquad \text{(A-A.3)} \end{aligned}$$

Where $\omega_i = -\beta_{ca} M_{error,i} + \nu_i$. Thus, the regression equation written in terms of M_i has an error term that contains the measurement error, $M_{error,i}$.

The probability limit of the OLS estimator for β_{ca} in Equation A-A.3 is:

$$\begin{aligned} plim(\hat{\beta}_{OLS}) &= \frac{COV(M_i, Y_i)}{VAR(M_i)} \\ &= \frac{COV(M_i, \alpha + \beta_{ca} M_i + \omega_i)}{VAR(M_i)} \\ &= \frac{COV(M_i, \alpha) + \beta_{ca} COV(M_i, M_i) + COV(M_i, \omega_i)}{VAR(M_i)} \end{aligned}$$

$$= \frac{0 + \beta_{ca} VAR(M_i) + COV(M_i, \omega_i)}{VAR(M_i)}$$

$$= \beta_{ca} + \frac{COV(M_i, \omega_i)}{VAR(M_i)} \qquad \text{(A-A.4)}$$

The assumptions of no omitted variable bias, no simultaneity bias, and purely random measurement error implies that:

$$COV(M_{true,i}, \nu_i) = 0 \qquad \text{(A-A.5)}$$

$$COV(M_{error,i}, \nu_i) = 0 \qquad \text{(A-A.6)}$$

$$COV(M_{true,i}, M_{error,i}) = 0 \qquad \text{(A-A.7)}$$

$$VAR(M_i) = VAR(M_{true,i}) + VAR(M_{error,i}) \qquad \text{(A-A.8)}$$

Combining Equations A-A.5-A-A.8 yields:

$$COV(M_i, M_{error,i}) = COV(\, M_{true,i} + M_{error,i},\ M_{error,i})$$

$$= COV(M_{true,i}, M_{error,i}) + COV(M_{error,i},\ M_{error,i})$$

$$= 0 + VAR(\, M_{error,i})$$

$$= VAR(M_{error,i}) \qquad \text{(A-A.9)}$$

Therefore:

$$COV(M_i, \omega_i) = COV(\, M_i, -\beta_{ca} M_{error,i} + \nu_i)$$

$$= -\beta_{ca} COV(\, M_i, M_{error,i}) + COV(M_i, \nu_i)$$

$$= -\beta_{ca} VAR(M_{error,i}) + COV(M_{true,i} + M_{error,i}, \nu_i)$$

$$= -\beta_{ca} VAR(M_{error,i}) + COV(M_{true,i}, \nu_i) + COV(M_{error,i}, \nu_i)$$

$$= -\beta_{ca} VAR(M_{error,i}) \qquad \text{(A-A.10)}$$

Substituting Equations A-A.8 and A-A.10 into Equation A-A.4 yields:

$$plim(\hat{\beta}_{OLS}) = \beta_{ca} + \frac{-\beta_{ca} VAR(M_{error,i})}{VAR(M_i)}$$

$$= \beta_{ca} \left(\frac{VAR(M_{true,i})}{VAR(M_{true,i}) + VAR(M_{error,i})}\right) \qquad \text{(A-A.11)}$$

Recall that λ is the reliability of the observed mediator, M_i, and:

$$\lambda \equiv \frac{VAR(M_{true,i})}{VAR(M_{true,i})+VAR(M_{error,i})} \qquad \text{(A-A.12)}$$

Therefore:

$$plim(\hat{\beta}_{OLS}) = \lambda\beta_{ca} \qquad \text{(A-A.13)}$$

Proof for Proposition A.2

The true relationship between the outcome and mediator is described by the following equation:

$$Y_i = \alpha + \beta_{ca}M_{true,i} + \beta_1 X_{1,i} + \nu_i \qquad \text{(A-A.14)}$$

where $X_{1,i}$ is a correctly measured covariate for individual i; β_{ca} is the causal relationship between the outcome and the mediator, conditioning on $X_{1,i}$; and $\beta_1$is the relationship between the outcome and the covariate $X_{1,i}$. All other variables are the same as defined in Equation A-A.1. Further assume that X_1 is uncorrelated with the measurement error for the mediator and the regression residuals.

Suppose that $M_{true,i}$ is measured imprecisely by M_i, that is:

$$M_i = M_{true,i} + M_{error,i} \qquad \text{(A-A.15)}$$

where the measurement error for individual i, $M_{error,i}$, is purely random with mean zero and variance $VAR(M_{error,i})$, and it is uncorrelated with $M_{true,i}$ and the regression error ν_i.

Note that because M_i, not $M_{true,i}$, is observed, the estimated regression equation is based on M_i. Substituting Equation A-A.15 into Equation A-A.14, therefore, yields:

$$Y_i = \alpha + \beta_{ca}M_i + \beta_1 X_{1,i} + \omega_i \qquad \text{(A-A.16)}$$

where $\omega_i = -\beta_{ca}M_{error,i} + \nu_i$. Consequently, the regression equation written in terms of M_i has an error term that contains measurement error, $M_{error,i}$. Let us denote the OLS estimator for β_{ca} as $\beta_{YM|X_1}$; that is, $\beta_{YM|X_1} \equiv \hat{\beta}_{OLS}$.

As a first step, note that $\beta_{YM|X_1}$ can be represented as follows:[27]

$$\beta_{YM|X_1} = \frac{\beta_{YM}-\beta_{YX_1}\beta_{MX_1}}{1-R^2_{MX_1}} \qquad \text{(A-A.17)}$$

[27] Blalock (1972), p. 451.

where $\beta_{YM}, \beta_{YX_1}, and\ \beta_{MX_1}$ are bivariate regression coefficients, and $R^2_{MX_1}$is the squared bivariate correlation between the observed mediator M and the covariate X_1. Note that $\beta_{YM|X_1}$ is the OLS estimator one would get from Equation A-A.16.

Using Equation A-A.17, we can adjust $\beta_{YM|X_1}$ for the reliability of M, λ (defined in Equation A-A.12), and thereby obtain an expression for the parameter we are interested in, $\beta_{ca|X_1}$. Specifically:

$$\beta_{ca|X_1} = \frac{\beta_{YM}/\lambda - \beta_{YX_1}(\beta_{MX_1}/\lambda)}{1 - R^2_{MX_1}/\lambda}$$

$$= \frac{\beta_{YM} - \beta_{YX_1}\beta_{MX_1}}{\lambda - R^2_{MX_1}} \qquad \text{(A-A.18)}$$

Combining Equations A-A.17 and A-A.18 yields:

$$\frac{\beta_{YM|X_1}}{\beta_{ca|X_1}} = \frac{\dfrac{\beta_{YM} - \beta_{YX_1}\beta_{MX_1}}{1 - R^2_{MX_1}}}{\dfrac{\beta_{YM} - \beta_{YX_1}\beta_{MX_1}}{\lambda - R^2_{MX_1}}}$$

$$= \frac{\lambda - R^2_{MX_1}}{1 - R^2_{MX_1}} \qquad \text{(A-A.19)}$$

Therefore:

$$\beta_{YM|X_1} = [\frac{\lambda - R^2_{MX_1}}{1 - R^2_{MX_1}}]\, \beta_{ca|X_1} \qquad \text{(A-A.20)}$$

In other words, the probability limit of the OLS estimator for M from Equation A-A.16, $\hat{\beta}_{OLS}$, is:

$$plim(\hat{\beta}_{OLS}) = [\frac{\lambda - R^2_{MX1}}{1 - R^2_{MX1}}]\beta_{ca} \qquad \text{(A-A.21)}$$

Appendix B

Assessment of a TSLS Estimator for a Single Mediator and Instrument in the Absence of Clustering

This appendix examines the bias-related statistical properties of a TSLS estimator of the causal relationship between an outcome and a single endogenous mediator using a single instrument. The discussion presented here focuses on the approximate properties of the *median* instead of the *expected value*, of the sampling distribution of the TSLS estimator, because for just-identified instrumental variable analysis (that is, when the number of instrumental variables is the same as the number of endogenous mediators), the expected value (or mean) of the TSLS estimator sampling distribution does not exist (Basman 1960, 1963; Bound, Jaeger, and Baker, 1995).[28] In addition, the appendix focuses on a situation where the mediator, the instrument, and the outcome are all measured at the unit level (for individuals or settings) — as opposed to the cluster level, and there is no cluster structure in the data. Furthermore, to simplify the discussion, there are no other exogenous covariates in the model. The appendix demonstrates the following:

- The median of a two-stage least-squares estimator for a single mediator and a single instrument in the absence of clustering can be approximately expressed as a weighted average of the true causal effect and the cross-sectional effect ***(Section II)***.
- In an experiment, median bias due to "weak instruments" in a two-stage least-squares estimator is approximately a proportion of existing cross-sectional bias (usually referred to as "OLS bias"), where the proportion is related to the amount of variation in the predicted value of a mediator that is "error-induced" versus "treatment-induced" ***(Section III)***.
- The first-stage population F-value, which the literature on instrumental variable methods has long advocated as a measure of the "strength" of an instrumental variable (for example, see Bound, Jaeger, and Baker, 1995), is approximately inversely proportional to the median bias in the two-stage least-squares estimator ***(Section IV)***.

One example of the situation assessed in this appendix is the "Moving to Opportunity" experiment (Kling, Liebman, and Katz, 2007), where individuals were randomly assigned to receive a housing voucher or not, the voucher was delivered to individuals directly, not through any cluster-level mediator (for example, not through schools or communities), and outcomes were measured for each individual. A similar example is an experiment that provides financial incentives to individual students. In other words, in this kind of situation, the mediator, the instrument, and the outcome are all measured at the unit level.

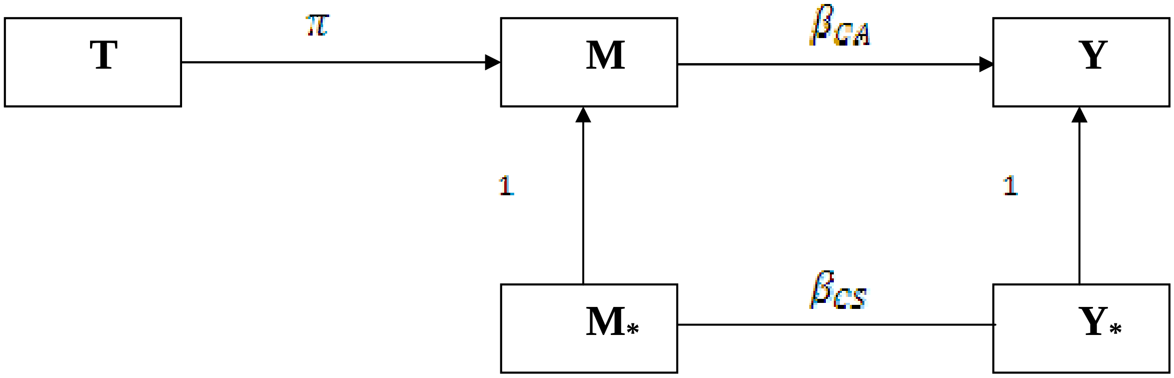
I. The Situation

Figure B.1 illustrates the situation being considered. It represents a series of relationships among a treatment indicator, T, a mediator, M, and an outcome, Y, with treatment status randomly assigned to individual sample members, i. The bottom two boxes in the diagram illustrate the underlying cross-

[28] For overidentified instrumental variable analysis (that is, when there are more instrumental variables than endogenous mediators in the model), the expected value of the TSLS estimator does exist.

Figure B.1

Cross-Section Bias from Weak Instruments



sectional relationship that exists between the outcome in the absence of treatment, Y_*, and the mediator in the absence of treatment, M_*.[29] This cross-sectional relationship reflects factors like attenuation bias and omitted variables bias in addition to any causal path that might exist between the two variables. The cross-sectional relationship can be modeled as:

$$Y_{*i}{=}\alpha_{cs} + \beta_{cs}M_{*i} + v_i \quad \text{(B.1)}$$

Figure B.1 also illustrates a causal effect, π, of treatment on the mediator, such that:

$$M_i{=}M_{i*} + \pi T_i \quad \text{(B.2)}$$

and a causal effect, $\pi\beta_{ca}$, of treatment on the outcome, such that:

$$Y_i = Y_{*i} + \pi\beta_{ca}T_i \quad \text{(B.3)}$$

Substituting Equation B.1 in to Equation B.3 yields:

$$Y_i = \alpha_{cs} + \beta_{cs}M_{*i} + \pi\beta_{ca}T_i{+}v_i \quad \text{(B.4)}$$

The resulting equation demonstrates that the observed value of the outcome for a given unit, Y_i, is a linear function of the value of the mediator for that unit in the absence of treatment, M_{*i} (with regression coefficient, β_{cs}) and whether the unit was randomized to treatment or control status, T_i (with regression

[29] M_{*i} and Y_{*i} cannot be observed for sample members who receive treatment.

coefficient, $\pi\beta_{ca}$). Equation B.4 also includes a random error,v_i, which is independently and identically distributed. This error is uncorrelated with both M_{*i} and T_i (a fact which becomes important later).

Given the situation illustrated by Figure B.1, two-stage least squares can be used to estimate the causal effect of M on Y from the following model:

First stage

$$M_i = \mu + \pi T_i + \varepsilon_i \qquad \text{(B.5)}$$

Second stage

$$Y_i = \alpha + \beta_{ca} M_i + v_i \qquad \text{(B.6)}$$

Using OLS to estimate parameters μ and π from the first-stage equation, predicted values of the mediator can be constructed as:

$$\widehat{M}_i = \hat{\mu} + \hat{\pi} T_i \qquad \text{(B.7)}$$

For the discussion below, the *estimated* effect of treatment on the mediator, $\hat{\pi}$, from the first-stage equation can be represented as the combination of its true value, π, and estimation error, ε_π. Similarly, the estimated intercept for the first-stage regression, $\hat{\mu}$, reflects the true value of the intercept and estimation error, ε_μ. That is:

$$\hat{\pi} = \pi + \varepsilon_\pi \qquad \text{(B.8)}$$

$$\hat{\mu} = \mu + \varepsilon_\mu \qquad \text{(B.9)}$$

Therefore, predicted values of the mediator can be represented as:

$$\widehat{M}_i = (\mu + \varepsilon_\mu) + (\pi + \varepsilon_\pi) T_i \qquad \text{(B.10)}$$

Note that here the main concern is the estimation error for the treatment, ε_π, which reflects the treatment and control group “mismatch” on counterfactual values of the mediator. Specifically:

$$\varepsilon_\pi = \overline{M}_{*T} - \overline{M}_{*C} \qquad \text{(B.11)}$$

The predicted values of the mediator are then substituted for observed values of M_i in the second-stage, Equation B.6, which is estimated using OLS with an adjustment of the standard errors to account for the fact that predicted values of the mediator were used instead of its actual values.[30] This second-stage regression produces the TSLS estimator, $\hat{\beta}_{TSLS}$.

[30] See Greene (1997, p. 295, p. 742) for a discussion of this adjustment.

II. The Approximate Median Value of the TSLS Estimator

This section derives the approximate median value of the TSLS estimator. Note that Equation B.4 indicates that the actual value of an individual outcome is a linear function of the counterfactual value of its mediator, times the cross-sectional coefficient, β_{cs}, plus the causal effect of the treatment on the outcome, $\pi\beta_{ca}$, for treatment group members or plus zero for control group members. Hence, systematic variation in the outcome reflects both the causal effect of treatment and the underlying cross-sectional coefficient.

Equation B.4 also implies that the difference in mean individual outcomes $(\bar{Y}_T - \bar{Y}_C)$ for a treatment group and control group — the estimated effect of treatment on the outcome — is:

$$\bar{Y}_T - \bar{Y}_C = \frac{\Delta Y}{\Delta T} = \beta_{cs}(\bar{M}_{*T} - \bar{M}_{*C}) + \pi\beta_{ca} + (\bar{v}_T - \bar{v}_C) \qquad \text{(B.12)}$$

Equations B.10 and B.11 imply that the difference in mean mediator values $(\bar{M}_T - \bar{M}_C)$ for a treatment group and control group — the estimated effect of treatment on the mediator — is:

$$\bar{M}_T - \bar{M}_C = \frac{\Delta M}{\Delta T} = (\bar{M}_{*T} - \bar{M}_{*C}) + \pi \qquad \text{(B.13)}$$

Hence, the Wald estimator (and its TSLS equivalent) of the effect of the mediator on the outcome is:

$$\hat{\beta}_{TSLS} = \frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{\beta_{cs}(\bar{M}_{*T}-\bar{M}_{*C})+\pi\beta_{ca}+(\bar{v}_T-\bar{v}_C)}{(\bar{M}_{*T}-\bar{M}_{*C})+\pi} \qquad \text{(B.14)}$$

Equation B.14 illustrates that in finite samples (the only type to which researchers have access) the estimator is an amalgam of the true causal coefficient, β_{ca} , and the cross-sectional coefficient, β_{cs}. It should be no surprise then, that as shown below, the expected value of the TSLS or Wald estimator is approximately a weighted average of the two coefficients.

The two-stage least-squares estimator of the causal relationship between the outcome and the mediator, $\hat{\beta}_{TSLS}$, is:

$$\hat{\beta}_{TSLS} = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(\hat{M}_i-\hat{\bar{M}})}{\sum_{i=1}^{N}(\hat{M}_i-\hat{\bar{M}})^2} \qquad \text{(B.15)}$$

where $\bar{Y}$ and $\hat{\bar{M}}$ are the grand mean of the outcome measure (Y_i) and the predicted value of the mediator ($\hat{M}_i$).

Proposition B.1: The median value of a TSLS or Wald estimator for the situation described in equations B.1-B.11 is approximately:

$$MEDIAN(\hat{\beta}_{TSLS}) \approx \frac{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}+[\sigma_{M*}^2]\beta_{cs}}{\sigma_{M*}^2+N\bar{T}(1-\bar{T})\pi^2} \qquad \text{(B.16)}$$

Demonstration of this proposition is attached at the end of this appendix.

Further note that the expression in the denominator in Equation B.16 demonstrates that the variation in the predicted value of the mediator ($\widehat{M_\iota}$) comes from two sources: the part that is induced by the treatment (treatment-induced variation, or *tiv*) and the part that is induced by the first-stage estimation error (error-induced variation, or *eiv*). The expected values of these two parts are defined as the following:

$$TIV \equiv E\{tiv\} = N\bar{T}(1-\bar{T})\pi^2 \quad \text{(B.17)}$$

$$EIV \equiv E\{eiv\} = \sigma_{M*}^2 \quad \text{(B.18)}$$

Substituting Equations B.17 and B.18 into Equation B.16 yields:

$$MEDIAN\{\hat{\beta}_{TSLS}\} \approx \frac{(TIV)\beta_{ca} + (EIV)\beta_{cs}}{TIV + EIV}$$

$$= \frac{(TIV + EIV)\beta_{ca} + (EIV)\beta_{cs} - (EIV)\beta_{ca}}{TIV + EIV}$$

$$= \beta_{ca} + \frac{EIV}{TIV+EIV}(\beta_{cs} - \beta_{ca}) \text{(B.19)}$$

A. Bias in the Two-Stage Least-Squares Estimation

This section derives the expression for the median bias in the TSLS estimator. Note that:

$$MEDIAN\ BIAS_{TSLS} \equiv \ MEDIAN\{\hat{\beta}_{TSLS}\} - \beta_{ca} \quad \text{(B.20)}$$

Substituting Equation B.20 into Equation B.19 yields:

$$MEDIANBIAS_{TSLS} \ \approx \frac{EIV}{TIV+EIV}(\beta_{cs} - \beta_{ca}) \quad \text{(B.21)}$$

Also note that:

$$BIAS_{OLS} \equiv \ \beta_{cs} - \beta_{ca} \quad \text{(B.22)}$$

Therefore:

$$MEDIANBIAS_{TSLS} \approx \frac{EIV}{TIV+EIV} BIAS_{OLS} \quad \text{(B.23a)}$$

In other words, the median bias in the two-stage least-squares estimator is a fraction of the corresponding OLS bias, where the fraction is the ratio of the error-induced variation to the total variation in predicted values of the mediator.

In addition, because OLS estimates are typically normally distributed, median OLS bias equals mean OLS bias and therefore:

$$MEDIANBIAS_{TSLS} \approx \left[\frac{EIV}{TIV+EIV}\right] MEDIANBIAS_{OLS} \qquad \text{(B.23b)}$$

B. The First-Stage F Statistic

A sample F statistic is used typically as a joint test of the null hypothesis that coefficients for *all* instruments in the first-stage regression are zero. In general:

$$F_{sample} = \frac{SS_p/df_p}{SS_E/df_E} \qquad \text{(B.24)}$$

where SS_p *and* SS_E are the sum of squares predicted by the first-stage regression and the sum of squared residual errors, respectively, and df_p *and* df_E are the degrees of freedom for the sum of squares predicted by the regression (L-1, L = number of instruments plus intercept) and degrees of freedom for the sum of squared residual errors (N-L, N = total number of observations. Applying these definitions to Equation B.24 yields:

$$F_{sample} = \frac{SS(\hat{M}_i)/(L-1)}{SS(\hat{\varepsilon}_i)/(N-L)} \qquad \text{(B.25)}$$

Proposition B.2: The first-stage population F-value for the situation described in equations B.1-B.11 is approximately:

$$F_{pop}^{(1)} \approx \frac{\sigma_{M*}^2 + N\bar{T}(1-\bar{T})\pi^2}{\sigma_{M*}^2} \qquad \text{(B.26)}$$

Note that $F_{pop}^{(1)}$ stands for the population first-stage F-value with a single instrument.

Substituting Equations B.17 and B.18 into Equation B.26 yields:

$$F_{pop}^{(1)} \approx \frac{TIV+EIV}{EIV} \qquad \text{(B.27)}$$

By substituting Equation B.27 into Equation B.23, the median bias of the TSLS estimator can be expressed in terms of the first-stage F-value and the OLS bias:

$$MEDIANBIAS_{TSLS} \approx \frac{1}{F_{pop}^{(1)}} BIAS_{OLS} \qquad \text{(B.28)}$$

C. Assessing Properties of Median Bias

So far in this appendix, we have provided an approximation of the median value of the TSLS estimator, derived the median bias using this approximation, and demonstrated how the median bias is linked to the OLS bias through the first-stage F-value. Here we provide both theoretical and empirical evidence from Monte Carlo simulations to show that the properties of the median bias we have derived and discussed so far trace the pattern of biases that we would likely see in real data.

Nelson and Startz (1990) provide a theoretical study of the properties of median bias for the TLSL estimator with a single mediator and instrument. They derive the exact small sample distribution of this estimator and provide proofs for properties of its median bias, which indicate that:

(1) it always lies somewhere between 0 and the corresponding OLS bias[31] and
(2) it approaches OLS bias as the instrument becomes weaker.[32]

In what follows, we present results from several sets of simulations that demonstrate these two properties of median bias. In addition, they illustrate that our expression for the TSLS median bias (Equation B.28) approximates these properties for a single instrument and mediator. We first present a graphical summary of simulations that demonstrate that the median value of the just-identified TSLS estimator approaches the OLS distribution as the instrument becomes weaker. We then present a summary of simulation results to demonstrate that the relationship between the strength of the instrument (as measured by the F-value of its first-stage regression) and median bias (as expressed in Equation B.28) generally holds for a variety of situations. These simulation results are consistent with the theoretical findings presented by Nelson and Startz (1990) and with our approximation for median bias.

Simulation Set-Up

The set-up for the simulation exercise is based on that in Angrist and Pischke (2008) but is modified to reflect the situation discussed in the present paper (in which individuals' treatment status is used as an instrument for an endogenous mediator). For ease of demonstration and without loss of generality, it is assumed that there are no other covariates in the model. Specifically, data are simulated using the following TSLS regression model:

First stage

$$M_i = \mu + \pi T_i + \varepsilon_i \qquad \text{(B.29)}$$

Second stage

$$Y_i = \alpha + \beta_{ca} M_i + \nu_i \qquad \text{(B.30)}$$

where:

$Y_i, M_i\ and\ T_i$ = values of the outcome, endogenous mediator, and treatment status;[33] for individual I;

[31] Here OLS bias is defined as the difference between the expected value of the estimated OLS coefficient and the true causal relationship between the outcome and the mediator. Note, though, that given the normal distribution of the OLS estimator in this case, the OLS bias is also equal to the difference between the median value of the estimated OLS coefficient and the true value of the coefficient.

[32] See Corollary 3.1 and 3.2 in Nelson and Startz (1990) for proofs of these two properties.

β_{ca}= true causal relationship between the outcome and the mediator, assumed to be 1 for all cases;

π = relationship between the mediator and the treatment status;

ν_i = error term in the second-stage regression, assumed to be independently and normally distributed with mean zero and unit variance; and

ε_i = error term in the first-stage regression. Different distributions are used to simulate this variable (discussed further below).

To generate a mediator M_i that is endogenous to the outcome, the underlying cross-sectional relationship between Y_i *and* M_i ("β_{cs}", as shown in the paper) is assumed to be 1.6 for all cases.[34] Recall Equation 36 from the main body of the paper that:

$$BIAS_{OLS} = \beta_{cs} - \beta_{ca} \qquad \text{(36 restated)}$$

Therefore, the OLS bias in the simulated results is fixed at 0.6. This value will serve as a benchmark for the TSLS median biases generated under different scenarios.

Graphical Presentation of the Relationship between TSLS Median Bias and Instrument Strength

Figure B.2 shows the Monte Carlo cumulative distribution functions (CDF) of the OLS estimator (solid line) and three TSLS estimators. These three TSLS estimators are generated using the set-up described above with instruments of different strength as measured by their relationship with the mediator, π, and their first-stage F-values: strong (π = 0.19, F = 10, long-dashed line), weaker (π = 0.077, F = 2.5, short-dashed line), and weakest (π = 0.001, F = 1, dotted line). The CDFs for the OLS estimator and the three TSLS estimators are based on 10,000 replications and assume that the distribution of the first-stage error term, ε_i, is standard normal and the T/C ratio is 1:1. All simulated datasets have a fixed sample size of 1,000.

As seen in Figure B.2, because of the correlation between error terms in the first and second stages, the OLS estimator is biased and centered at a value of about 1.60. The TSLS estimator based on a strong instrument is centered around 1.00, and is virtually median unbiased. *The TSLS estimator using a weaker instrument is centered at 1.17, and the one with the weakest instrument* is centered at 1.60.

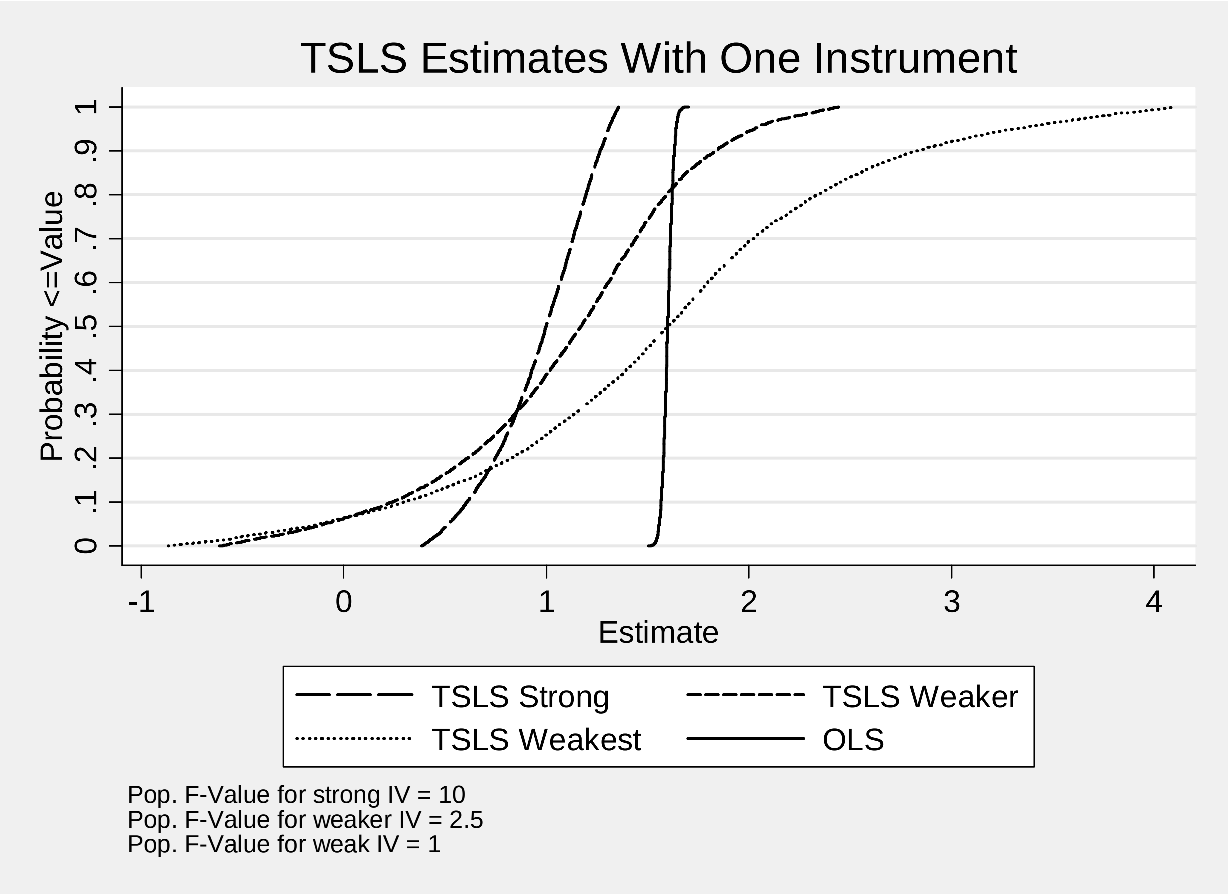
The simulation results shown in this figure visually demonstrate the two properties of median bias for a just-identified TSLS estimator discussed in Nelson and Startz (1990): (1) when the instrument is strong, the TSLS estimator is approximately median unbiased, and (2) as the instrument gets weaker, the

[33] Different values of this coefficient are simulated to reflect different levels of strength of the instrument.

[34] This is accomplished by constraining the correlation between the two error terms in the regressions, ν_i and ε_i, to be 0.6.

Figure B.2

Graphical Presentation of the Relationship between TSLS Median Bias and Instrument Strength



NOTE: For each case, extreme estimates (smaller/greater than the 10th/90th percentiles) are excluded from the graph.

distribution for a just-identified 2SLS estimator shifts toward the distribution of the OLS estimator, and therefore the median of the TSLS estimator moves further away from the true value and toward the OLS estimator. The median of the TSLS estimator eventually coincides with that of the OLS estimator when the instrument has no explanatory power for the mediator (F-value = 1).[35]

[35] In the tables, sometimes the estimated median for TSLS estimators exceeds the OLS estimator, this is because these simulation results are based on 10,000 replications and might not fully characterize the true distribution of the TSLS estimator.

Table B.1

Median Bias of the Just-Identified 2SLS Estimator, First- and Second-Stage Error Terms Jointly Normally Distributed

Panel A: T/C=1				
π	Population F-Value	Median TSLS Estimate	Median Bias…	
			Estimated	Implied by F-Value
0.001	1	1.6	0.6	0.6
0.045	1.5	1.38	0.38	0.4
0.063	2	1.26	0.26	0.3
0.077	2.5	1.17	0.17	0.24
0.089	3	1.11	0.11	0.2
0.126	5	1.02	0.02	0.12
0.190	10	1	0	0.06
0.237	15	1	0	0.04
Panel B: T/C=3				
0.001	1	1.59	0.59	0.6
0.052	1.5	1.37	0.37	0.4
0.073	2	1.24	0.24	0.3
0.089	2.5	1.17	0.17	0.24
0.103	3	1.12	0.12	0.2
0.146	5	1.03	0.03	0.12
0.219	10	1.01	0.01	0.06
0.273	15	1	0	0.04

Generalizability of TSLS Median Bias Properties

Tables B.1-B.4 present simulation results that expand on Figure B.2 in two ways: (1) in addition to the simulation of the first- and second-stage regression error terms using joint standard normal distributions (Table B.1), the first-stage regression error term (ε_i) is also generated using other widely used distributions that one would likely see, such as uniform distribution (Table B.2), log-normal distribution (Table B.3), and Gamma distribution (Table B.4);[36] and (2) in addition to examining situations where there are equal numbers of treatment and control group members (T/C ratio =1), we also extend the results to unbalanced designs. Specifically, we examine cases in which the T/C ratio is 3.

These tables have two panels and five columns. Panel A in each table shows the simulation results for the balanced design, and panel B shows the results for cases with a T/C ratio of 3.

[36] The second-stage regression error term is always generated using a standard normal distribution, because in education studies, the outcome measure is most likely to be student achievement measured by test scores, which are usually normally distributed.

Table B.2

Median Bias of the Just-Identified 2SLS Estimator, First- and Second-Stage Error Uniformly and Normally Distributed

Panel A: T/C=1				
π	Population F-Value	Median TSLS Estimate	Median Bias…	
			Estimated	Implied by F-Value
0.001	1	1.59	0.59	0.6
0.045	1.5	1.36	0.36	0.4
0.063	2	1.22	0.22	0.3
0.077	2.5	1.14	0.14	0.24
0.089	3	1.09	0.09	0.2
0.126	5	1.01	0.01	0.12
0.190	10	0.99	0.01	0.06
0.237	15	0.99	0.01	0.04
Panel B: T/C=3				
0.001	1	1.62	0.62	0.6
0.052	1.5	1.38	0.38	0.4
0.073	2	1.24	0.24	0.3
0.089	2.5	1.15	0.15	0.24
0.103	3	1.1	0.1	0.2
0.146	5	1.02	0.02	0.12
0.219	10	1	0	0.06
0.273	15	1	0	0.04

The first two columns of each table present specific values of key parameters used in the simulations. Column 1 reports the value of the regression coefficient for treatment in the first-stage regression (π). Column 2 shows the F-value for the first-stage regression. Recall Equation 43 in the paper:

$$F_{pop}^{(1)} = \frac{TIV+EIV}{EIV} = 1 + \frac{TIV}{EIV}$$

$$= 1 + \frac{N\bar{T}(1-\bar{T})\pi^2}{\sigma_{M*}^2} = 1 + N\bar{T}(1-\bar{T})(\frac{\pi}{\sigma_{M*}})^2 \qquad \text{(43 restated)}$$

$$= 1 + N\bar{T}(1-\bar{T})(ES_M)^2$$

It shows that the first-stage population F-value can be simulated based on three key parameters: the sample size, N, which is fixed at 1,000; the proportion of treatment observations, $\bar{T}$, which is fixed at 0.5 for the first panel and at 0.75 for the second panel in all tables; and the effect size of treatment on the mediator, ES_M, which is equivalent to $\frac{\pi}{\sigma_{M*}}$. Note that in all simulations, the variance of the counterfactual mediator is assumed to be 1. As a result, $ES_M = \frac{\pi}{\sigma_{M*}} = \frac{\pi}{1} = \pi$, which are the values reported in Column 1.

Table B.3

Median Bias of the Just-Identified 2SLS Estimator, First- and Second-Stage Error Terms Log-Normally and Normally Distributed

Panel A: T/C=1				
π	Population F-Value	Median TSLS Estimate	Median Bias…	
			Estimated	Implied by F-Value
0.001	1	1.61	0.61	0.6
0.045	1.5	1.39	0.39	0.4
0.063	2	1.24	0.24	0.3
0.077	2.5	1.17	0.17	0.24
0.089	3	1.11	0.11	0.2
0.126	5	1.02	0.02	0.12
0.190	10	1	0	0.06
0.237	15	1	0	0.04
Panel B: T/C=3				
0.001	1	1.61	0.61	0.6
0.052	1.5	1.38	0.38	0.4
0.073	2	1.25	0.25	0.3
0.089	2.5	1.16	0.16	0.24
0.103	3	1.11	0.11	0.2
0.146	5	1.03	0.03	0.12
0.219	10	1	0	0.06
0.273	15	1	0	0.04

The next two columns of each table report the simulation results. Column 3 shows the median of the estimated TSLS coefficient for the mediator based on the 10,000 replications. The next column shows the median bias values that are calculated as the difference between the median estimator and the true causal relationship between the outcome and the mediator (= 1 for all cases).

The last column in each table provides the TSLS bias as implied by the F-values reported in column 2. Values reported in this column are calculated by using Equation B.28:

$$MEDIANBIAS_{TSLS} \approx \frac{1}{F_{pop}^{(1)}} BIAS_{OLS} \tag{B.28 restated}$$

As discussed in the simulation set-up section, the OLS bias is fixed at 0.6 for all simulations. Therefore, the implied TSLS bias is completely determined by the population first-stage F-value. By comparing values reported in column 4, which come from the numerical simulations, with the numbers reported in column 5, which are based on analytical approximations derived and presented in the paper, one can assess the performance and accuracy of the intuition for finite sample bias presented in the paper.

Table B.4

Median Bias of the Just-Identified 2SLS Estimator, First- and Second-Stage Error Terms Gamma and Normally Distributed

Panel A: T/C=1				
π	Population F-Value	Median TSLS Estimate	Median Bias…	
			Estimated	Implied by F-Value
0.001	1	1.6	0.6	0.6
0.045	1.5	1.38	0.38	0.4
0.063	2	1.25	0.25	0.3
0.077	2.5	1.17	0.17	0.24
0.089	3	1.12	0.12	0.2
0.126	5	1.03	0.03	0.12
0.190	10	1.01	0.01	0.06
0.237	15	1	0	0.04
Panel B: T/C=3				
0.001	1	1.6	0.6	0.6
0.052	1.5	1.36	0.36	0.4
0.073	2	1.24	0.24	0.3
0.089	2.5	1.16	0.16	0.24
0.103	3	1.1	0.1	0.2
0.146	5	1.03	0.03	0.12
0.219	10	1	0	0.06
0.273	15	1	0	0.04

These four tables present simulated results for a wide range of different scenarios, and common patterns can be seen across different scenarios generated by varying the distribution of the variable of interest, data structure, and the strength of the instrument. In particular, the following patterns emerge from these tables:

1. The median bias of the TSLS estimator is inversely related to the strength of the instrumental variable. The stronger the instrument, the smaller the median bias is. When the instrument is very strong (with F-values equal to or greater than 10, as suggested by the conventional wisdom), the median bias approaches zero; when the instrument is very weak (with F-values at the minimum possible value of 1), the median bias approaches the OLS bias of 0.6; and

2. The median bias calculated based on simulations approximately traces the bias implied by the formula derived in the paper, indicating that the theoretical analysis provides a useful approximation for the pattern of biases that we would likely see in real data. This is especially true when the instrument is relatively weak (with F-values smaller than 10). Also note that the simulated median bias is always smaller than the one implied by the theoretical calculation, hence the latter is a conservative assessment of potential bias in the TSLS estimator.

Demonstrations of Propositions B.1 and B.2

Demonstration of Proposition B.1

The TSLS estimator of the relationship between an outcome and a mediator, $\hat{\beta}_{TSLS}$, is:

$$\hat{\beta}_{TSLS} = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(\widehat{M}_\mathrm{i}-\bar{\widehat{M}})}{\sum_{i=1}^{N}(\widehat{M}_\mathrm{i}-\bar{\widehat{M}})^2} \qquad \text{(A-B.1)}$$

where $\bar{Y}$ and $\bar{\widehat{M}}$ are the grand means of the outcome measure (Y_i) and the predicted value of the mediator ($\widehat{M_\iota}$). We first derive the expected value of the numerator of this expression and then derive the expected value of its denominator.

Numerator of the Estimator

Given Equation B.4, the first term in the numerator of Equation A-B.1 is:

$$Y_i - \bar{Y} = (\alpha_{cs} + \beta_{cs}M_{*i}+\pi\beta_{ca}T_i + \nu_i) - (\alpha_{cs} + \beta_{cs}\bar{M}_*+\pi\beta_{ca}\bar{T}+\bar{\nu})$$

$$= \beta_{cs}(M_{*i} - \bar{M}_*) \ + \ \pi\beta_{ca}(T_i - \bar{T}) + (\nu_i - \bar{\nu})$$

$$\text{(A-B.2)}$$

where $\bar{M}_*$ is the sample mean value of M_* and $\bar{T}$ is the sample mean value of T_i (which equals the proportion of sample members randomized to treatment, $\bar{T}$).

The second term in the numerator of Equation A-B.1 is:

$$\widehat{M}_\mathrm{i} - \bar{\widehat{M}} = [\hat{\mu} + (\pi + \epsilon_\pi)T_i] - [\hat{\mu} + (\pi + \varepsilon_\pi)\bar{T}]$$

$$= (\pi + \varepsilon_\pi)(T_i - \bar{T}) \qquad \text{(A-B.3)}$$

Substituting Equations A-B.2 and A-B.3 into the numerator of Equation A-B.1 and taking the expected value of the result yields:

$$E\{\textstyle\sum_{i=1}^{N}[\beta_{cs}(M_{*i} - \bar{M}_*) + \pi\beta_{ca}(T_i - \bar{T}) + (\nu_i - \bar{\nu})][(\pi + \varepsilon_\pi)(T_i - \bar{T})]\} \qquad \text{(A-B.4)}$$

To help keep track of the next several steps it is useful to consider separately the following three components of Equation A-B.4.

Component #1

$$E\{\textstyle\sum_{i=1}^{N}[\beta_{cs}(M_{*i} - \bar{M}_*)](\pi + \varepsilon_\pi)(T_i - \bar{T})\}$$

$$= \beta_{cs}E\{(\pi + \varepsilon_\pi)\textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\}$$

$$= \pi\beta_{cs}E\{\textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\} + \beta_{cs}E\{\varepsilon_\pi \textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\}$$

$$= 0 + \beta_{CS}\mathrm{E}\{\varepsilon_\pi \textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\}$$

$$= \beta_{cs}E\{\varepsilon_\pi \textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\} \qquad \text{(A-B.5)}$$

Note that the expected value of the first summation in the third line of Equation A-B.5 equals zero, because randomization ensures that, in expectation, treatment status is uncorrelated with any pre-existing characteristic of sample members.

To proceed further, note that estimation error, ε_π, for the first-stage regression coefficient is

$$\varepsilon_\pi = \bar{M}_{*T} - \bar{M}_{*C} \qquad \text{(A-B.6)}$$

where $\overline{M_{*T}}\ and\ \overline{M_{*C}}$ are treatment and control group mean values of the mediator in the absence of treatment.[37] Substituting this fact into Equation A-B.5 yields:

$$\beta_{cs}E\{\varepsilon_\pi \sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\}$$

$$= \beta_{cs}E\{(\overline{M_{*T}} - \overline{M_{*C}}) \textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T})\} \qquad \text{(A-B.7)}$$

The next steps require decomposing the summation in Equation A-B.7 into its counterparts for the treatment group and control group as follows:

$$\textstyle\sum_{i=1}^{N}(M_{*i} - \bar{M}_*)(T_i - \bar{T}) = \sum_{i=1}^{N\bar{T}}(M_{*i} - \bar{M}_*)(1 - \bar{T}) + \sum_{i=N\bar{T}+1}^{N}(M_{*i} - \bar{M}_*)(0 - \bar{T})$$

Treatment group *Control Group*

$$= (1 - \bar{T})\textstyle\sum_{i=1}^{N\bar{T}}(M_{*i} - \bar{M}_*) + (0 - \bar{T})\sum_{i=N\bar{T}+1}^{N}(M_{*i} - \bar{M}_*)$$

$$= (1 - \bar{T})N\bar{T}(\bar{M}_{*T} - \bar{M}_*) - \bar{T}N(1 - \bar{T})(\bar{M}_{*C} - \bar{M}_*)$$

$$= N\bar{T}(1 - \bar{T})(\bar{M}_{*T} - \bar{M}_{*C}) + N\bar{T}(1 - \bar{T})(-\overline{M_*} + \bar{M}_*)$$

$$= N\bar{T}(1 - \bar{T})(\bar{M}_{*T} - \bar{M}_{*C}) + 0$$

$$= N\bar{T}(1 - \bar{T})(\bar{M}_{*T} - \bar{M}_{*C}) \qquad \text{(A-B.8)}$$

[37] As noted, the observed treatment-group and control-group difference of mean values for M equals the true impact of treatment on the mediator, π, plus estimation error, ε_π. Estimation error in this case equals the treatment group and control group "mismatch" with respect to the mean value of the mediator in absence of treatment. This mismatch is the result of "imperfect draws" that occur by chance in finite samples.

Substituting Equation A-B.8 into Equation A-B.7 yields:

$$\beta_{cs}\mathrm{E}\{\varepsilon_\pi \textstyle\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{M}_{*\mathrm{i}} - \bar{M}_*)(\mathrm{T_i} - \bar{\mathrm{T}})\} = \beta_{\mathrm{cs}}\mathrm{E}\{(\bar{M}_{*T} - \bar{M}_{*C})N\bar{T}(1-\bar{T})(\bar{M}_{*T} - \bar{M}_{*C})\}$$

$$= \beta_{cs}N\bar{T}(1-\bar{T})E\{(\bar{M}_{*T} - \bar{M}_{*C})^2\}$$

$$= \beta_{cs}N\bar{T}(1\text{-}\bar{T})VAR(\bar{M}_{*T} - \bar{M}_{*C})$$

$$= \beta_{cs}N\bar{T}(1\text{-}\bar{T})(\frac{\sigma^2_{M*}}{N\bar{T}(1-\bar{T})})$$

$$= \boldsymbol{\beta_{cs}\sigma^2_{M_*}} \qquad \text{(A-B.9)}$$

Component #2

Next, note that:[38]

$$\pi\beta_{ca}E\{(\pi + \varepsilon_\pi)\textstyle\sum_{i=1}^{N}(T_i - \bar{T})^2\} = \pi\beta_{ca}E\{(\pi + \varepsilon_\pi)\, N\bar{T}(1-\bar{T})\}$$

$$= [N\bar{T}(1-\bar{T})\pi^2]\beta_{ca} + [N\bar{T}(1-\bar{T})\pi\beta_{ca}]E\{\varepsilon_\pi\}$$

$$= [N\bar{T}(1-\bar{T})\pi^2]\beta_{ca} + 0$$

$$= \boldsymbol{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}} \qquad \text{(A-B.10)}$$

Component #3

Lastly, note that:

$$E\{\textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(\pi + \varepsilon_\pi)(T_i - \bar{T})\} = E\{(\pi + \varepsilon_\pi)\textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\}$$

$$= \pi E\{\textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\} + E\{\varepsilon_\pi \textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\}$$

$$= 0 + E\{\varepsilon_\pi \textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\}$$

$$= E\{(\bar{M}_{*T} - \bar{M}_{*C})\textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\} \qquad \text{(A-B.11)}$$

Again, it is useful to separate the components of the summation into those representing the treatment group and those representing the control group. Doing so yields:

$$E\{(\bar{M}_{*T} - \bar{M}_{*C})\textstyle\sum_{i=1}^{N}(\nu_i - \bar{\nu})(T_i - \bar{T})\}$$

$$= E\{(\bar{M}_{*T} - \bar{M}_{*C})[\textstyle\sum_{i=1}^{N\bar{T}}(\nu_i - \bar{\nu})(1-\bar{T}) + \textstyle\sum_{i=N\bar{T}+1}^{N}(\nu_i - \bar{\nu})(0-\bar{T})]\}$$

[38] The variance of a dichotomous variable, T, equals $\bar{T}(1-\bar{T})$. Hence, its total variation in a sample of size N is $N\bar{T}(1-\bar{T})$.

$$= E\{(\bar{M}_{*T} - \bar{M}_{*C})[(1 - \bar{T})\sum_{i=1}^{N\bar{T}}(\nu_i - \bar{\nu}) + (0 - \bar{T})\sum_{i=N\bar{T}+1}^{N}(\nu_i - \bar{\nu})]\}$$

$$= E\{(\bar{M}_{*T} - \bar{M}_{*C})[N\bar{T}(1 - \bar{T})(\bar{\nu}_T - \bar{\nu}) - N\bar{T}(1 - \bar{T})(\bar{\nu}_C - \bar{\nu})]\}$$

$$= N\bar{T}(1 - \bar{T})E\{(\bar{M}_{*T} - \bar{M}_{*C})[(\bar{\nu}_T - \bar{\nu}_C) + (-\bar{\nu} + \bar{\nu})]\}$$

$$= N\bar{T}(1 - \bar{T})E\{(\bar{M}_{*T} - \bar{M}_{*C})(\bar{\nu}_T - \bar{\nu}_C)\} \qquad \text{(A-B.12)}$$

As noted earlier, when describing Equation B.4, ν is uncorrelated with M*. Therefore:

$$N\bar{T}(1 - \bar{T})E\{(\bar{M}_{*T} - \bar{M}_{*C})(\bar{\nu}_T - \bar{\nu}_C)\} = N\bar{T}(1 - \bar{T})(0)$$

$$= \mathbf{0} \qquad \text{(A-B.13)}$$

Combining the Components

The expected value of the numerator of the TSLS estimator therefore is:

$$E\{\sum_{i=1}^{N}(Y_i - \bar{Y})\left(\widehat{M_\iota} - \bar{\widehat{M}}\right) = \beta_{cs}\sigma^2_{M_*} + [N\bar{T}(1 - \bar{T})\pi^2]\beta_{ca} + 0$$

$$= (\boldsymbol{\sigma}^2_{\boldsymbol{M*}})\boldsymbol{\beta_{cs}} + [\boldsymbol{N\bar{T}(1 - \bar{T})\pi^2}]\boldsymbol{\beta_{ca}} \qquad \text{(A-B.14)}$$

This result is a weighted sum of the cross-sectional regression coefficient, β_{cs}, and the causal regression coefficient, β_{ca}.

Denominator of the Estimator

Now consider the expected value of the denominator of the TSLS estimator, where:

$$E\left\{\sum_{i=1}^{N}(\widehat{\mathrm{M_i}} - \bar{\widehat{M}})^2\right\} = E\left\{\sum_{i=1}^{N}\{[\hat{\mu} + (\pi + \epsilon_\pi)T_i]\text{- } [\hat{\mu} + (\pi + \varepsilon_\pi)\bar{T}]\}^2\right\}$$

$$= E\{\sum_{i=1}^{N}[(\pi + \varepsilon_\pi)(T_i - \bar{T})]^2\}$$

$$= E\{(\pi + \varepsilon_\pi)^2\sum_{i=1}^{N}[(T_i - \bar{T})]^2\}$$

$$= E\{(\pi + \varepsilon_\pi)^2 N\bar{T}(1 - \bar{T})\}$$

$$= N\bar{T}(1 - \bar{T})E\{(\pi + \varepsilon_\pi)^2\}$$

$$= N\bar{T}(1 - \bar{T})E\{\pi^2 + 2\pi\varepsilon_\pi + {\varepsilon_\pi}^2\}$$

$$= N\bar{T}(1 - \bar{T})\{E(\pi^2) + E(2\pi\varepsilon_\pi) + E({\varepsilon_\pi}^2)\}$$

$$= N\bar{T}(1 - \bar{T})\{\pi^2 + 2\pi E(\varepsilon_\pi) + E({\varepsilon_\pi}^2)\}$$

$$= N\bar{T}(1 - \bar{T})\{\pi^2 + 0 + E({\varepsilon_\pi}^2)\}$$

$$= N\bar{T}(1-\bar{T})\{\pi^2 + VAR(\varepsilon_\pi)\} \qquad \text{(A-B.15)}$$

$VAR(\varepsilon_\pi)$ is simply $VAR(\hat{\pi})$, the variance of estimation error for the first-stage regression coefficient, where:

$$VAR(\hat{\pi}) = \frac{\sigma^2_{M_*}}{N\bar{T}(1-\bar{T})} \qquad \text{(A-B.16)}$$

Substituting Equation A-B.16 into Equation A-B.15 yields:

$$E\left\{\sum_{i=1}^{N}(\widehat{M}_{\mathrm{i}} - \bar{\widehat{M}})^2\right\} = N\bar{T}(1-\bar{T})\pi^2 + \sigma^2_{M_*} \qquad \text{(A-B.17)}$$

The Full Expression

Combining the preceding findings and rearranging terms yields:

$$\frac{E\{\sum_{i=1}^{N}(Y_i-\bar{Y})(\widehat{M}_{\mathrm{i}}-\bar{\widehat{M}})\}}{E\{\sum_{i=1}^{N}(\widehat{M}_{\mathrm{i}}-\bar{\widehat{M}})^2\}} = \frac{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}+[\sigma^2_{M*}]\beta_{cs}}{\sigma^2_{M*}+N\bar{T}(1-\bar{T})\pi^2} \qquad \text{(A-B.18)}$$

nominator in Equation A-B.18 is close to a constant across sampling replications (that is, its variance is small). Therefore, it is expected that the TSLS estimator is approximately equal to:[39]

$$E(\,\hat{\beta}_{TSLS}\,) \approx \frac{E\{\sum_{i=1}^{N}(Y_i-\bar{Y})(\widehat{M}_{\mathrm{i}}-\bar{\widehat{M}})\}}{E\{\sum_{i=1}^{N}(\widehat{M}_{\mathrm{i}}-\bar{\widehat{M}})^2\}}$$

$$= \frac{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}+[\sigma^2_{M*}]\beta_{cs}}{\sigma^2_{M*}+N\bar{T}(1-\bar{T})\pi^2} \qquad \text{(A-B.19)}$$

On the other hand, for a just-identified instrumental variable analysis, the expected value of the TSLS estimator does not exist. In other words, in theory, the expected value of the TSLS estimator does not exist for a single mediator and single instrument case. However, the median of this estimator does exist, and the literature has been assessing the properties of the TSLS estimator for the just-identified cases through the median of the estimator distribution instead (for example, see Angrist and Pischeke, 2008, among others). Following this tradition, we express the median of the TSLS estimator using the expression derived in equation A-B.18:

$$MEDIAN(\hat{\beta}_{TSLS}) \approx \frac{[N\bar{T}(1-\bar{T})\pi^2]\beta_{ca}+[\sigma^2_{M*}]\beta_{cs}}{\sigma^2_{M*}+N\bar{T}(1-\bar{T})\pi^2} \qquad \text{(A-B.20)}$$

Even though we do not provide a direct proof of this expression, theoretical and simulated empirical evidence provided above demonstrate that the properties of the TSLS median bias derived using this expression (especially how the median bias is linked to the OLS bias through the first-stage F-value)

[39] A similar approximation was used by Hahn and Hausman (2002).

provide a useful approximation for the pattern of biases that we would likely see in real data. These evidences indicate that the approximation of the median value for the TSLS estimator expressed in equation A-B.20 is valid.

Proof of Proposition B.2

The situation considered in this appendix follows the set-up laid out by Equations B.1-B.11. Furthermore, for the first-stage regression in a TSLS analysis, Equation A-B.17, shows that the expected value of the sum of squares predicted by the regression can be expressed as the following:

$$E(SS_p) = E(SS(\widehat{M}_{\mathrm{i}})) = E\{\sum_{i=1}^{N}\left(\widehat{M}_{\mathrm{i}} - \bar{\widehat{M}}\right)^2\}$$

$$= \sigma^2_{M*} + N\bar{T}(1-\bar{T})\pi^2 \qquad \text{(A-B.21)}$$

Similarly, the expected value of the sum of squared residual errors is:

$$E(SS_E) = E(SS(\widehat{\varepsilon}_{\mathrm{i}})) = E\{\sum_{i=1}^{N}\left(M_i - \widehat{M}_{\mathrm{i}}\right)^2\}$$

$$= (N-L)var(\varepsilon_i)$$

$$= (N-L)\sigma^2_{M*} \qquad \text{(A-B.22)}$$

Even though $\sigma^2_{M*}$is random, as $N \rightarrow \infty$, asymptotically, we expect the population F-value to approximately equal:

$$F_{pop} = E\left(F_{sample}\right) = E\left(\frac{SS(\widehat{M}_i)/(L-1)}{SS(\hat{\varepsilon}_i))/(N-L)}\right) \approx \frac{E(SS(\widehat{M}_i))/(L-1)}{E(SS(\hat{\varepsilon}_i))/(N-L)}$$

$$= \frac{(\sigma^2_{M*} + N\bar{T}(1-\bar{T})\pi^2)/(L-1)}{(N-L)\sigma^2_{M*}/(N-L)}$$

$$= \frac{(\sigma^2_{M*} + N\bar{T}(1-\bar{T})\pi^2)/(L-1)}{\sigma^2_{M*}} \qquad \text{(A-B.23)}$$

Note that this approximation rests on the fact that sample-based estimates of a population variance are quite accurate (they have little sampling variability) if they are based on more than about 20 degrees of freedom. This point can be illustrated by the relationship that exists between a t distribution and a normal or z distribution. A t-statistic is the ratio of a sample-based parameter estimate to the sample-based estimate of its standard deviation (the square root of its variance). A z-statistic has the same numerator but assumes that the standard deviation (and thus variance) of the parameter is known. When the standard deviation of the estimator is estimated with very few degrees of freedom, the critical value for a t distribution (say for a two-tail hypothesis test at the 0.05 level of statistical significance) is much larger than that for a z distribution. This reflects the uncertainty — and thus variability — that exists for a sample-based estimate of a standard deviation or variance given very few degrees of freedom. For example, with only four degrees of freedom, the 0.05 two-tail critical value is 2.78 for a t-statistic versus 1.96 for a z-statistic. As the number of degrees of freedom (and thus sample size) increases, the critical value of a t-statistic rapidly approaches that of a z-statistic. For example, with 20 degrees of freedom the 0.05 two-tail critical value of a t-statistic is 2.09.

Because in the current appendix we consider the case of one instrumental variable, L-1 = 1. It follows that:

$$F_{POP}^{(1)} \approx \frac{\sigma_{M*}^2 + N\bar{T}(1-\bar{T})\pi^2}{\sigma_{M*}^2} \qquad \text{(A-B.24)}$$

where $F_{POP}^{(1)}$ stands for the population first-stage F-value for a single instrument.

Appendix C

Assessment of a TSLS Estimator for a Single Mediator and Instrument for Clustered Samples

This appendix assesses the bias of a TSLS estimator and its F-value for a single mediator and a single instrument for clustered samples. Similar to Appendix B, we focus on the approximate statistical properties of the median of the sampling distribution of the just-identified TSLS estimator rather than its expected value, which does not exist. In other words, in what follows, we assess the finite sample bias of the just-identified TSLS estimator in clustered data structures through the median value of its sampling distribution.

Recall from the main body of the paper that we consider two different prototypical situations: individual-level analysis and setting-level analysis. In the first situation, the mediator and outcome vary across individuals. For example, the mediator might be individual student engagement, and the outcome might be individual student achievement; individual student being the unit of analysis. In this situation, individuals are clustered if they are randomized in groups and/or they are treated in groups; hence the instrument (treatment status) varies only by group (or cluster).

In the second situation, the mediator is a setting-level characteristic, and the outcome is either inherently a setting-level characteristic or is an individual-level characteristic that is aggregated to the setting level, usually by averaging. For example, the setting mediator might be a specific classroom instructional practice and the setting outcome might be average student achievement for each classroom. In this situation, settings are clustered if they are randomized and/or treated in clusters; thus the instrument varies by cluster.

The convention that we use for both individual-level situations and setting-level situations is as follows. We refer to settings or individuals as *units* and to interdependent groups of *units* that are randomized and/or treated together as clusters. Specifically, we consider a situation where there are $\underline{J}$ clusters with a constant number of $\underline{n}$ units per cluster. Clusters are randomized in proportion $\bar{T}$ to the treatment group and $(1-\bar{T})$ to the control group.

In this appendix, we study the statistical properties of a TSLS or Wald estimator of the causal relationship between an outcome and mediator in the unit-level data through those in corresponding aggregate cluster-level data.[40] For simplicity, there are no other exogenous covariates in the model. We summarize our main results as follows:

- The median of the sampling distribution of the TSLS estimator in the presence of clustering can be approximately expressed as a weighted average of the true causal effect and the underlying cross-sectional effect. ***(Section III)***

[40] Recall that in a setting-level analysis, the setting-level characteristics can themselves be aggregates of individual characteristics. For such cases, the expression "aggregate cluster-level data" refers to the data that is constructed by further aggregating setting-level characteristics to the higher-level of clusters, by which settings are randomized and/or treated.

- Finite sample median bias of the TSLS estimator in the presence of clustering is a fraction of the "OLS bias," where the particular proportion equals the error-induced proportion of total variation in predicted values of the mediator. ***(Section IV)***

- The clustered first-stage F-statistic is inversely proportional to the bias in the median of the TSLS estimator. ***(Section V)***

- Other things being equal (including the total variation in counterfactual values of the mediator), clustering *increases* the error-induced variation in predicted values of the mediator and *decreases* the first-stage F-statistic; in this way, clustering *increases* finite sample bias in a TSLS estimator. ***(Sections III, IV, and V)***

I. Situation

For typical unit-level data with clustering introduced above, the first-stage and second-stage regressions become:

First stage

$$M_{ij} = \mu + \pi T_j + e_j + \varepsilon_{ij} \qquad \text{(C.1)}$$

Second stage

$$Y_{ij} = \alpha + \beta_{ca} M_{ij} + w_j + v_{ij} \qquad \text{(C.2)}$$

where:

M_{ij} = the mediator for unit *i* in cluster *j,*

T_j = the treatment status indicator for cluster *j*[41],

Y_{ij} = the outcome for unit *i* in cluster *j,*

e_j and ε_{ij} = the random error for cluster *j* and unit *i* in cluster *j*, respectively. These errors are assumed to be independent of each other and distributed with mean zero and variance of $\tau^2_{M_*}$ and $\theta^2_{M_*}$, respectively.

w_j and v_{ij} = the random error for cluster *j* and unit *i* in cluster *j*, respectively. These errors are assumed to be independent of each other and distributed with mean zero and variance of $\tau^2_{Y_*}$ and $\theta^2_{Y_*}$, respectively.

π = the effect of the treatment on the mediator,

[41] Note that this term does not carry the index for individuals, *i*, since random assignment takes place at the cluster level and thus treatment/control status of all individuals within a cluster is constant.

β_{ca} = the causal effect of the mediator on the outcome.

Note that we can express the relationship between unit-level and cluster-level variance components of the first-stage regression as an intra-class correlation, ρ. This parameter is defined as the ratio of the cluster-level variance component to the sum of the cluster-level and unit-level variance components:

$$\rho_{M*} = \frac{\tau^2_{M*}}{\tau^2_{M*} + \theta^2_{M*}} \tag{C.3}$$

In order to compare the statistical properties of TSLS estimators in the presence of clustering with those in the absence of clustering (Appendix B), it is necessary to hold constant the variance of the mediator. This implies that the total unit variance without clustering, σ^2_{M*}, equals the total unit variance with clustering, $\tau^2_{M*} + \theta^2_{M*}$. Using this condition:

$$\rho_{M*} = \frac{\tau^2_{M*}}{\sigma^2_{M*}} \tag{C.4}$$

and

$$1 - \rho_{M*} = \frac{\theta^2_{M*}}{\sigma^2_{M*}} \tag{C.5}$$

As described in the main body of the paper, TSLS estimation of the models depicted by Equations C.1 and C.2 proceeds as follows:

- The first-stage regression (Equation C.1) is estimated using OLS.
- OLS estimates of the intercept ($\hat{\mu}_U$) and coefficient ($\hat{\pi}_U$) from the first stage are used to predict the value of the mediator for each unit as:

$$\hat{M}_{ij} = \hat{\mu}_U + \hat{\pi}_U T_j \tag{C.6}$$

Note that the subscript "U" is referring to estimates based on unit-level data.

Since the value of the treatment indicator, T_j, in Equation C.6 is constant for all units in a given cluster, values of the predicted mediator for all those in that cluster are also constant ($\hat{M}_{ij} = \hat{M}_j$).

- Predicted values of the mediator, $\hat{M}_{ij}$, are substituted for actual values of M_{ij} in the second-stage regression (Equation C.2), whose intercept and coefficient are estimated using OLS.[42] The resulting estimate of causal coefficient β_{ca} (which represents the causal relationship between the outcome and the mediator) is the TSLS estimate, $\hat{\beta}_{TSLS(U)}$.

In the following sections we discuss how properties of $\hat{\beta}_{TSLS(U)}$ can be studied using the properties of the TSLS estimate of the causal coefficient obtained from aggregate data created for each cluster.

II. Using Properties of the Aggregate TSLS Estimator to Study the Properties of the Unit-Level TSLS Estimator

The simplest way to examine the situation summarized above is through the use of aggregate data in the TSLS estimation. The following proposition motivates our approach.

Proposition C.1: Consider a simple regression model that employs an outcome and independent variable. OLS estimates of the model's intercept and coefficient in a unit-level dataset are equivalent to those in the corresponding aggregate data when each aggregate has the same number of units, and the independent variable is constant across units within an aggregate entity.

Proof of this proposition is provided at the end of this appendix.

Utilizing aggregate data in TSLS estimation entails (i) creating cluster-level means of the dependent and independent variables in Equations C.1 and C.2 and (ii) using the cluster-level variables in the TSLS estimation as described above. Specifically, for each cluster, we have: T_j, which represents its treatment status; $\overline{Y}_j$, its mean value for the outcome; and $\overline{M}_j$, its mean value for the mediator. The first-stage regression for the aggregate data is then:

$$\overline{M}_j = \mu + \pi T_j + \overline{\varepsilon}_j \tag{C.7}$$

where $\overline{\varepsilon}_j$ is the mean error for cluster *j:*

$$\overline{\varepsilon}_j \equiv \frac{\sum_{i=1}^{n}[e_j + \varepsilon_{ij}]}{n} = e_j + \frac{\sum_{i=1}^{n}\varepsilon_{ij}}{n} \tag{C.8}$$

and is assumed to be distributed with mean zero and variance $\overline{\sigma}^2_{M*}$

[42] Note that it is the standard practice to adjust the standard errors of these OLS estimates to account for (i) the clustered structure of the data and (ii) the use of predicted values of the endogenous regressors in place of the actual ones.

As for the unit-level data, TSLS estimation using aggregate data starts with estimation of the first-stage regression (Equation C.7) by OLS. Let us denote the resulting estimates of the intercept and the treatment effect on the mediator as $\hat{\mu}_A$ and $\hat{\pi}_A$, respectively. The relationship between these and their unit-level counterparts is summarized in the following lemma:

Lemma C.1: OLS estimates of the intercept and the coefficient of the first-stage regression are equivalent for the unit-level and aggregate data.

$$\hat{\mu}_A = \hat{\mu}_U \text{ and } \hat{\pi}_{\mathrm{A}} = \hat{\pi}_{\mathrm{U}} \tag{C.9}$$

Proof. This result follows directly from Proposition C.1 as (i) each aggregate (or cluster) has the same number of units and (ii) the independent variable in used in both regressions is constant across units within an aggregate entity.

Next, the predicted values of aggregate mediator, $\hat{\overline{M}}_j$, are estimated using $\hat{\mu}_A$ and $\hat{\pi}_A$ as :

$$\hat{\overline{M}}_j = \hat{\mu}_A + \hat{\pi}_A T_j \tag{C.10}$$

As a side note, observe that the predicted mediator $\hat{\overline{M}}_j$ contains endogenous variation from random estimation error in $\hat{\pi}_A$, ε_π, which reflects the mismatch on $\overline{M}_j$ between the treatment and control group. Therefore:

$$\hat{\overline{M}}_j = \hat{\mu}_A + \hat{\pi}_A T_j = \hat{\mu}_A + (\pi + \varepsilon_\pi) T_j \tag{C.11}$$

Finally, the TSLS estimate of the causal coefficient β_{ca} in the aggregate data, $\hat{\beta}_{TSLS(A)}$, is calculated through the OLS analysis of the second-stage regression using the aggregate outcome, $\overline{Y}_j$, and the predicted aggregate mediator, $\hat{\overline{M}}_j$.

Lemma C.2: The TSLS estimates of the causal coefficient in the unit-level data and the aggregate data are equivalent.

$$\hat{\beta}_{TSLS(U)} = \hat{\beta}_{TSLS(A)} \tag{C.12}$$

Proof. Once again this result follows from Proposition C.1. Note that (i) the independent variable used in the estimation of the second-stage regression with unit-level data ($\hat{M}_{ij}$) is constant within an aggregate entity ($\hat{M}_{ij} = \hat{M}_j$) and (ii) its values are equal to those of the regressor used in the second-stage estimation with the aggregate data ($\hat{M}_{ij} = \hat{M}_j = \hat{\overline{M}}_j$), which can be verified by comparing Equations C.6 and C.10 using Equation C.9.

In the following sections, we provide an approximation for the median value of $\hat{\beta}_{TSLS(A)}$ and its corresponding population F-value, which are in turn used to analyze properties of $\hat{\beta}_{TSLS(U)}$ given Equation C.12. Utilizing the resemblance of the situation with aggregate data to the one with nonclustered data, these derivations draw heavily from the results established in the absence of clustering in Appendix B.

III. The Approximated Median Value of a TSLS Estimator for a Single Mediator and Instrument in the Presence of Clustering

First note that Equation B.4 takes the following form for aggregate data:

$$\overline{Y}_j = \alpha_{cs} + \beta_{cs}\overline{M}_{*j} + \pi\beta_{ca}T_j + \overline{v}_j \tag{C.13}$$

Equation C.13 demonstrates that the observed mean value of the outcome in a given cluster, $\overline{Y}_j$, is a linear function of the mean value of the mediator in that cluster in the absence of the treatment, $\overline{M}_{*j}$ (with the regression coefficient β_{cs}) and whether the cluster was randomized to treatment or control status, T_j (with regression coefficient $\pi\beta_{ca}$).[43] This regression also includes a cluster-level error term, $\overline{v}_j$, which is independently and identically distributed.

The TSLS estimate of the relationship between the outcome and the mediator for the aggregate data, $\hat{\beta}_{TSLS(A)}$, is:

$$\hat{\beta}_{TSLS(A)} = \frac{\sum_{j=1}^{J}(\overline{Y}_j - \overline{Y})(\hat{\overline{M}}_j - \hat{\overline{M}})}{\sum_{j=1}^{J}(\hat{\overline{M}}_j - \hat{\overline{M}})^2} \tag{C.14}$$

where $\overline{Y}$ and $\hat{\overline{M}}$ are sample mean values of the outcome and predicted mediator.[44]

Proposition C.2: The median value of the TSLS estimator in Equation C.14 is approximately:

$$Median\{\hat{\beta}_{TSLS(A)}^{(K)}\} \approx \frac{[nJ\overline{T}(1-\overline{T})\pi^2]\beta_{ca} + \sigma_{M*}^2[1+(n-1)\rho_{M*}]\beta_{cs}}{nJ\overline{T}(1-\overline{T})\pi^2 + \sigma_{M*}^2[1+(n-1)\rho_{M*}]} \tag{C.15}$$

Demonstration of this proposition is provided at the end of this appendix.

From Lemma C.2, it is easy to see that median value of the aggregate TSLS estimator is equal to that of the unit-level TSLS estimator, which is given by Equation C.15. Note that the typical-unit level

[43] In this appendix, we focus on the simple case in which the underlying cross-sectional relationship in the unit-level data is equivalent to that in the aggregate data.

[44] Note that sample mean values of the aggregate variables are the same as those of the unit-level variables, since we assume that each cluster has the same number of individuals.

data considered in this appendix has a cluster structure. Therefore, in order to distinguish results established here from those established in the absence of clustering, we use the subscript "CL" for the appropriate terms, starting with the unit-level TSLS estimator. Hence, using Proposition C.2 and Lemma C.2, we posit that:

$$MEDIAN\{\hat{\beta}_{TSLS(CL)}\} \equiv MEDIAN\{\hat{\beta}_{TSLS(U)}\} =$$

$$MEDIAN\{\hat{\beta}_{TSLS(A)}\} \approx \frac{[nJ\bar{T}(1-\bar{T})\pi^2]\beta_{ca} + \sigma^2_{M*}[1+(n-1)\rho_{M*}]\beta_{cs}}{nJ\bar{T}(1-\bar{T})\pi^2 + \sigma^2_{M*}[1+(n-1)\rho_{M*}]} \qquad \text{(C.16)}$$

As in Appendix B, we can express Equation C.16 in terms of the expected values of the treatment-induced variation, TIV_{CL}, and the error-induced variation, EIV_{CL}, in predicted values of the mediator in the presence of clustering. Let us define:

$$TIV_{CL} \equiv E\{tiv_{CL}\} = nJ\bar{T}(1-\bar{T})\pi^2 \qquad \text{(C.17)}$$

$$EIV_{CL} \equiv E\{eiv_{CL}\} = \sigma^2_{M*}[1+(n-1)\rho_{M*}] \qquad \text{(C.18)}$$

Note from Equations B.17 and B.18 that in the absence of clustering:

$$TIV \equiv E\{tiv\} = nJ\bar{T}(1-\bar{T})\pi^2 \qquad \text{(B.17 restated)}$$

$$EIV \equiv E\{eiv\} = \sigma^2_{M*} \qquad \text{(B.18 restated)}$$

Comparing Equations C.17 and B.17 suggests that clustering does not affect treatment-induced variation in predicted values of the mediator, that is, $TIV_{CL} = TIV$.

Comparing Equations C.18 and B.18, however, implies that clustering increases the error-induced variation in predicted values of the mediator by a factor of $[1+(n-1)\rho_{M*}]$.

Next, substituting Equations C.17 and C.18 in Equation C.16 yields:

$$MEDIAN\{\hat{\beta}_{TSLS(CL)}\} \approx \frac{TIV}{TIV + EIV_{CL}}\beta_{ca} + \frac{EIV_{CL}}{TIV + EIV_{CL}}\beta_{cs} \qquad \text{(C.19)}$$

Equation C.19 provides valuable insights into the effect of clustering on the median value of the TSLS estimate. Specifically, Equation C.19 in conjunction with Equation B.19 implies that, other things being equal, clustering reduces the relative weight placed by TSLS on the true causal effect (β_{ca}) of a mediator; thereby increasing the relative weight of the underlying cross-sectional coefficient (β_{cs}) since EIV_{CL} is greater than EIV.

IV. Median Bias of the TSLS Estimator in the Presence of Clustering

Recall that finite sample median bias for a two-stage least-squares estimator is defined as the difference between its median value and β_{ca}. Using this definition and Equation C.19:

$$MEDIANBIAS_{TSLS(CL)} \equiv MEDIAN\{\hat{\beta}_{TSLS(CL)}\} - \beta_{ca} \approx \frac{EIV_{CL}}{TIV + EIV_{CL}}[\beta_{cs} - \beta_{ca}] \qquad \text{(C.20)}$$

As in Appendix B, the difference between β_{cs} and β_{ca} can be referred to as "OLS bias." Using this definition, the relationship between finite sample bias for TSLS and OLS bias in the presence of clustering is:[45]

$$MEDIANBIAS_{TSLS(CL)} \approx \frac{EIV_{CL}}{TIV + EIV_{CL}} BIAS_{OLS} \qquad \text{(C.21)}$$

The effect of clustering on finite sample median bias of a TSLS estimator is summarized in the following proposition:

Proposition C.3: Other things being equal (such as total variance of the mediator and the total number of units), clustering increases the magnitude of median bias in TSLS estimators:

$$\left|MEDIANBIAS_{TSLS(CL)}\right| > \left|MEDIANBIAS_{TSLS}\right| \qquad \text{(C.22)}$$

Proof of this proposition is provided at the end of the appendix.

V. F-Value of a TSLS Estimator for a Single Mediator and Instrument in the Presence of Clustering

As in Appendix B, we first derive the population F-statistic for the first-stage regression in the cluster-level aggregate data. Proposition C.4 summarizes the resulting expression:

Proposition C.4: The first-stage population F-value for aggregate data is approximately:

$$F^{(1)}_{pop(A)} \approx \frac{nJ\overline{T}(1-\overline{T})\pi^2 + \sigma^2_{M_*}[1+(n-1)\rho_{M_*}]}{\sigma^2_{M_*}[1+(n-1)\rho_{M_*}]} \qquad \text{(C.23)}$$

Proof of this proposition is provided at the end of the appendix.

Note that this statistic can be used as the F-value for the unit-level data with clustering, since estimation of the aggregate model and the unit-level model produces identical results (Lemmas 1 and 2 in the main text). Therefore:

[45] As stated in Appendix B, since OLS estimates are typically normally distributed, their mean and median values are equal. Hence, the bias of the OLS estimate can also be expressed in terms of its median value.

$$F^{(1)}_{pop(CL)} \equiv F^{(1)}_{pop(U)} = F^{(1)}_{pop(A)} \approx \frac{nJ\overline{T}(1-\overline{T})\pi^2 + \sigma^2_{M_*}[1+(n-1)\rho_{M_*}]}{\sigma^2_{M_*}[1+(n-1)\rho_{M_*}]} \tag{C.24}$$

Once again, the F-statistic represents the inverse of the ratio of error-induced variation to total variation in predicted values of the mediator. Specifically, substituting Equations C.17 and C.18 in Equation C.24, we obtain:

$$F^{(1)}_{pop(CL)} \approx \frac{TIV + EIV_{CL}}{EIV_{CL}} \tag{C.25}$$

Also note that using Equation C.25 in conjunction with Equation C.20 yields:

$$MEDIANBIAS_{TSLS(CL)} \approx \frac{1}{F^{(1)}_{pop(CL)}} BIAS_{OLS} \tag{C.26}$$

Using Equation C.25 in conjunction with Equation B.27, it is easy to see that clustering reduces the population F-value, thereby reducing the strength of the instrument. Equation C.26 also suggests that clustering increases the median bias of the TSLS estimator, a result already shown in the previous section.

Proofs of Propositions C.1-C.4

Proof of Proposition C.1

Consider a typical situation in which units (for example, students) are clustered within J aggregate clusters (for example, schools), each with n units. We specify the following regression model to represent the relationship between an unit-level outcome, Y_{ij}, and a cluster-level independent variable, X_j, where subscripts *i* and *j* represent units and clusters, respectively:

$$Y_{ij} = A + BX_j + c_j + d_{ij} \qquad \text{(A-C.1)}$$

In this model, c_j and d_{ij} are the random error terms for clusters and units, respectively, and are assumed to be independent of each other. Also note that the value of X_j is constant for all units in cluster *j*.

Next, consider the aggregate (cluster-level) data which is constructed using cluster-level means of the outcome ($\overline{Y}_j$) and the independent variable (X_j, since it is constant within clusters). The corresponding aggregate regression model is then:

$$\overline{Y}_j = A + BX_j + \overline{c}_j \qquad \text{(A-C.2)}$$

where $\overline{c}_j$ is the cluster-level error term. We posit that the OLS estimate of the coefficient B using unit-level data, $\hat{B}_U$, is equivalent to the one yielded using aggregate data, $\hat{B}_A$. A proof of this statement is as follows:

$$\begin{aligned}
\hat{B}_U &= \frac{COV(Y_{ij}, X_j)}{VAR(X_j)} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}(Y_{ij} - \overline{Y})(X_j - \overline{X})}{\sum_{j=1}^{J}\sum_{i=1}^{n}(X_j - \overline{X})^2} \\
&= \frac{\sum_{j=1}^{J}(X_j - \overline{X})\sum_{i=1}^{n}(Y_{ij} - \overline{Y})}{n\sum_{j=1}^{J}(X_j - \overline{X})^2} \\
&= \frac{\sum_{j=1}^{J}(X_j - \overline{X})n(\overline{Y}_j - \overline{Y})}{n\sum_{j=1}^{J}(X_j - \overline{X})^2} \\
&= \frac{\sum_{j=1}^{J}(X_j - \overline{X})n(\overline{Y}_j - \overline{Y})}{n\sum_{j=1}^{J}(X_j - \overline{X})^2} \\
&= \frac{\sum_{j=1}^{J}(X_j - \overline{X})(\overline{Y}_j - \overline{Y})}{\sum_{j=1}^{J}(X_j - \overline{X})^2} = \frac{COV(\overline{Y}_j, X_j)}{VAR(X_j)} = \hat{B}_A
\end{aligned} \qquad \text{(A-C.3)}$$

where $\overline{Y}$ and $\overline{X}$ are sample mean values of the outcome and the independent variable, respectively and they are equal in the unit-level and aggregate data. Note that OLS estimates of the intercept of the unit-level and aggregate regressions are also equivalent:

$$\begin{aligned} \hat{A}_U &= \overline{Y} - \hat{B}_U \overline{X} \\ &= \overline{Y} - \hat{B}_A \overline{X} \\ &= \hat{A}_A \end{aligned} \qquad \text{(A-C.4)}$$

Demonstration of Proposition C.2

As mentioned in the main text, aggregate cluster-level data considered in this appendix are essentially identical to the typical unit-level data without clustering which are analyzed in Appendix B.[46] Hence results established in Appendix B for the median bias of the TSLS estimator in the absence of clustering can be directly applied to the situation considered here. Specifically, adapting Equation B.16 to the current case yields the approximate median value of the aggregate TSLS estimator:

$$MEDIAN\{\hat{\beta}_{TSLS(A)}\} \approx \frac{[J\overline{T}(1-\overline{T})\pi^2]\beta_{ca} + \overline{\sigma}^2_{M*}\beta_{cs}}{J\overline{T}(1-\overline{T})\pi^2 + \overline{\sigma}^2_{M*}} \qquad \text{(A-C.5)}$$

where $\overline{\sigma}^2_{M*}$ is the variance of the aggregate error term from the first-stage regression in Equation C.7, which replaced the term σ^2_{M*} in Equation B.16. We also use *J* in Equation A-C.5 in place of *N* in Equation B.16 since the aggregate data has J observations (or clusters).

Using Equation C.8, $\overline{\sigma}^2_{M*}$ can be expressed in terms of the underlying variance structures in the unit-level data as:

$$\begin{aligned} \overline{\sigma}^2_{M*} &= VAR(\overline{\varepsilon}_j) = VAR\left[e_j + \frac{\sum_{i-1}^{n} \varepsilon_{ij}}{n} \right] = VAR(e_j) + \frac{1}{n^2}\sum_{i=1}^{n} VAR(\varepsilon_{ij}) \\ &= VAR(e_j) + \frac{1}{n^2} nVAR(\varepsilon_{ij}) = \tau^2_{M*} + \frac{\theta^2_{M*}}{n} \end{aligned} \qquad \text{(A-C.6)}$$

where we use the fact that e_j and ε_{ij} are independent and $\text{cov}(\varepsilon_{kj}, \varepsilon_{lj})$ =0 for any *k* and *l* by construction. Further note that using Equations C.4 and C.5, we can express the cluster-level and unit-level variance

[46] One can compare the first- and second-stage regressions in the aggregate data with those in the unit-level data without clustering to see the validity of this statement. Note that although error terms in the aggregate models inherently represent the unit- and cluster-level variance structures of the underlying unit-level data, there is no explicit clustering in the aggregate data since it contains only one observation for each aggregate entity.

components in Equation A-C.6 in terms of the total variance of units within and between clusters, σ^2_{M*}, and the intra-class correlation, ρ_{M*}. That is:

$$\tau^2_{M*} = \rho_{M*}\sigma^2_{M*} \text{ and } \theta^2_{M*} = (1-\rho_{M*})\sigma^2_{M*} \qquad \text{(A-C.7)}$$

Substituting Equation A-C.7 in Equation A-C.6 yields:

$$\overline{\sigma}^2_{M*} = \rho_{M*}\sigma^2_{M*} + \frac{(1-\rho_{M*})\sigma^2_{M*}}{n} = \frac{\sigma^2_{M*}}{n}[1+(n-1)\rho_{M*}] \qquad \text{(A-C.8)}$$

Finally, substituting Equation A-C.8 in A-C.5 and multiplying the numerator and the denominator by n yields:

$$MEDIAN\{\hat{\beta}_{TSLS(A)}\} \approx \frac{[nJ\overline{T}(1-\overline{T})\pi^2]\beta_{ca} + \sigma^2_{M*}[1+(n-1)\rho_{M*}]\beta_{cs}}{nJ\overline{T}(1-\overline{T})\pi^2 + \sigma^2_{M*}[1+(n-1)\rho_{M*}]} \qquad \text{(A-C.9)}$$

We can also derive the variance of the aggregate (or clustered) first-stage coefficient estimate using Equation A-C.8. Note that from Lemma C.1 and adapting Equation A-B.16 for the present case, we get:

$$\begin{aligned} VAR(\hat{\pi}_{CL}) \equiv VAR(\hat{\pi}_U) = VAR(\hat{\pi}_A) &= \frac{\overline{\sigma}^2_{M*}}{J\overline{T}(1-\overline{T})} \\ &= \frac{\sigma^2_{M*}}{\overline{T}(1-\overline{T})}\left[\frac{\rho_{M*}}{J} + \frac{1-\rho_{M*}}{nJ}\right] \end{aligned} \qquad \text{(A-C.10)}$$

Proof of Proposition C.3

Note that median bias of the TSLS estimator in the absence and presence of clustering was expressed as:

$$MEDIANBIAS_{TSLS} \approx \frac{EIV}{TIV + EIV} BIAS_{OLS} \qquad \text{(B.23 restated)}$$

$$MEDIANBIAS_{TSLS(CL)} \approx \frac{EIV_{CL}}{TIV + EIV_{CL}} BIAS_{OLS} \qquad \text{(C.21 restated)}$$

The ratio of Equation C.21 to B.23 is:

$$\frac{MEDIANBIAS_{TSLS(CL)}}{MEDIANBIAS_{TSLS}} = \frac{\dfrac{EIV_{CL}}{TIV + EIV_{CL}}}{\dfrac{EIV}{TIV + EIV}} = \frac{EIV_{CL} \cdot TIV + EIV_{CL} \cdot EIV}{EIV \cdot TIV + EIV \cdot EIV_{CL}}$$

$$= \frac{EIV_{CL} \cdot EIV}{EIV_{CL} \cdot EIV} \frac{1+\dfrac{TIV}{EIV}}{1+\dfrac{TIV}{EIV_{CL}}} = \frac{1+\dfrac{TIV}{EIV}}{1+\dfrac{TIV}{EIV_{CL}}} > 1$$

$$\to \left|MEDIANBIAS_{TSLS(CL)}\right| > \left|MEDIANBIAS_{TSLS}\right| \qquad \text{(A-C.11)}$$

where we utilize the fact that $EIV_{CL} > EIV$.

Proof of Proposition C.4

As in the proof of Proposition C.2, we can use the results established in Appendix B in the absence of clustering for the derivation of the first-stage F-statistic in the aggregate data when a single instrument and mediator are used. Specifically, adapting Equation B.26 (replacing $\sigma^2_{M_*}$ with $\bar{\sigma}^2_{M_*}$ and N with J) for the current case yields:

$$F^{(1)}_{pop(A)} \approx \frac{J\bar{T}(1-\bar{T})\pi^2 + \bar{\sigma}^2_{M_*}}{\bar{\sigma}^2_{M_*}} \qquad \text{(A-C.12)}$$

Substituting the expression for $\bar{\sigma}^2_{M_*}$ from A-C.8 in A-C.12 and multiplying the numerator and the denominator by n yields:

$$F^{(1)}_{pop(A)} \approx \frac{nJ\bar{T}(1-\bar{T})\pi^2 + \sigma^2_{M_*}[1+(n-1)\rho_{M_*}]}{\sigma^2_{M_*}[1+(n-1)\rho_{M_*}]} \qquad \text{(A-C.13)}$$

Appendix D

Finite Sample Bias with a Single Mediator and Multiple Randomized Instruments

In this appendix, we study the statistical properties of an overidentified TSLS estimator that uses a single mediator and multiple instruments that are created from treatment indicators for multiple studies, sites, or subgroups — referred to as "strata." Unlike the single instrument and single mediator case considered in Appendixes B and C, the expected value of the overidentified TSLS estimator exists. Moreover, since the distribution of the overidentified TSLS estimator is asymptotically normal, its asymptotic mean equals its asymptotic median. Therefore, we assess its finite sample bias through both its mean and median.

In what follows, we consider two situations: (i) the general case where true treatment effects on the mediator vary across strata (according to a prespecified rule) and (ii) the more specific case where treatment effects are constant across strata. For each situation, we also compare results produced by multiple instruments with those produced by a single instrument to determine if and when using multiple instruments is worthwhile.

The appendix proceeds as follows:

- Given a set of assumptions described in ***Section I***, the expected value and variance of the treatment effect on mediator in stratum K, π_k, are derived in ***Section II***.
- Treatment-induced and error-induced variations in predicted values of the mediator are derived and compared with their counterparts for a single instrument (***Section III***).
- The expected value and bias of the two-stage least-squares estimator under current conditions is derived (***Sections IV and V***).
- Conditions are derived for the strength of the set of K instruments to exceed that of a single instrument, and therefore reduce finite sample bias (***Section VI***).

I. Situation

Suppose in a randomized experiment, units (individuals or settings) are randomly assigned to the treatment or control group separately in K study strata. The conceptual model of the first and second stages of a TSLS analysis with treatment status indicator T, mediator, M, and outcome Y is then:

$$M_{ik} = \mu_k + \pi_k T_{ik} + \varepsilon_{ik} \tag{D.1}$$

$$Y_{ik} = \alpha_k + \beta M_{ik} + \upsilon_{ik} \tag{D.2}$$

where i represents unit i and k represents stratum k. Predicted values of the mediator (which are substituted for their actual values in the second-stage regression) are then:

$$\hat{M}_{ik} = \hat{\mu}_k + \hat{\pi}_k T_{ik} \tag{D.3}$$

For simplicity, we assume that there is no clustering and that each of the K strata has the same number of units, proportion of units randomized to treatment, and counterfactual variance of the mediator. That is:

Assumption D.1: $N_1 = N_2 = \cdots = N_K = \frac{1}{K}N,$

where N is the total sample size, K is the total number of strata, and N_k $(k = 1, 2, 3 \ldots K)$ is the sample size for the kth stratum.

Assumption D.2: $\bar{T}_1 = \bar{T}_2 = \cdots = \bar{T}_K = \bar{T},$

where $\bar{T}$ is the proportion of units randomized to treatment for the full sample and $\bar{T}_k$ $(k = 1, 2, 3 \ldots K)$ is its counterpart for the kth stratum

Assumption D.3: $\sigma^2_{M_*(1)} = \sigma^2_{M_*(2)} = \cdots = \sigma^2_{M_*(K)} = \sigma^2_{M_*},$

where $\sigma^2_{M_*}$is the variance of counterfactual values of the mediator for the full sample and $\sigma^2_{M_*(k)}$ $(k = 1, 2, 3 \ldots K)$ is its counterpart for the kth stratum.

More importantly, also assume, for now, that the true impact of treatment on the mediator varies by site but follows the following rules: If sites are ordered by the true impact of treatment on the mediator, with site 1 having the least positive impact and site K having the most positive impact, the true treatment effect on the mediator is:

Assumption D.4: $\pi_k = \pi_{k-1} + \phi,$

where $k = 1, 2, 3 \ldots K$, with π_1 as the smallest treatment effect on the mediator, and ϕ a nonnegative constant.

In what follows, the discussion will focus on the general case where ϕ is nonnegative (henceforth referred to as "general case"). We will also present results for the special case of $\phi = 0$, which represents the situation where the true treatment effect on the mediator, π_k , is constant across K strata (henceforward referred to as "special case"). For this special case, therefore, we denote $\pi_k = \pi$ for all k.

II. Mean and Variance of π_k

Proposition D.1: Based on Assumptions D.1-D.4, it can be shown that the mean and variance of π_k are:

$$E(\pi_k) = \pi_1 + \frac{K-1}{2}\phi \qquad \text{(D.4)}$$

$$VAR(\pi_k) = \phi^2(\frac{K^2-1}{12}) \qquad \text{(D.5)}$$

Proof for this proposition can be found at the end of this appendix.

The following table illustrates how the mean and variance of π_k varies with the number of sites, K. Note that we use $\phi_{(K)}$to denote the nonnegative constant in each scenario.

Number of Sites (K)	1	2	3	4	10	20
$\boldsymbol{E(\pi_k)}$	π_1	$\pi_1 + \frac{1}{2}\phi_{(2)}$	$\pi_1 + \phi_{(3)}$	$\pi_1 + \frac{3}{2}\phi_{(4)}$	$\pi_1 + \frac{9}{2}\phi_{(10)}$	$\pi_1 + \frac{19}{2}\phi_{(20)}$
$\boldsymbol{VAR(\pi_k)}$	0	$\frac{1}{4}\phi_{(2)}^2$	$\frac{2}{3}\phi_{(3)}^2$	$\frac{5}{4}\phi_{(4)}^2$	$\frac{33}{4}\phi_{(10)}^2$	$\frac{133}{4}\phi_{(20)}^2$

Note that for the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k$,

$E(\pi_k) = \pi$

$VAR(\pi_k) = 0$

III. Expectations of Treatment-Induced Variation (TIV) and Error-Induced Variation (EIV) for K Instruments

Using Equations D.4 and D.5, and adapting Equation B.17, it can be shown that the expectation of the treatment-induced variation for 1 instrument in a first-stage regression (for a sample with K sites), $E(tiv^{(1)})$, is:

$$E(tiv^{(1)}) = N\bar{T}(1-\bar{T})[E(\pi_k)]^2$$

$$= N\bar{T}(1-\bar{T})(\pi_1 + \tfrac{K-1}{2}\phi)^2 \qquad \text{(D.6)}$$

Proposition D.2: The expectation of treatment-induced variation for K instruments in a first-stage regression (for a sample with K sites), $E(tiv^{(K)})$, is:

$$TIV^{(K)} \equiv E(tiv^{(K)}) = N\bar{T}(1-\bar{T})\{\pi_1{}^2 + (K-1)\pi_1\phi + \phi^2\tfrac{(K-1)(2K-1)}{6}\} \qquad \text{(D.7)}$$

Proof for this proposition is provided at the end of this appendix.

For the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k$,

$$E(tiv^{(K)}) = N\bar{T}(1-\bar{T})[E(\pi_k)]^2$$

$$= N\bar{T}(1-\bar{T})(\pi)^2$$

Based on Equation B.18, it can also be shown that

$$EIV^{(K)} \equiv E\left(eiv^{(K)}\right) = N\bar{T}(1-\bar{T})[VAR\left(\hat{\varepsilon}_{\pi}^{(K)}\right)]$$

$$= N\bar{T}(1-\bar{T})\frac{\sigma_{M_*}^2}{\frac{N}{K}\bar{T}(1-\bar{T})}$$

$$= K\sigma_{M_*}^2 \qquad \text{(D.8)}$$

For the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k$,

$$E\left(eiv^{(K)}\right) = K\sigma_{M_*}^2$$

IV. Expected Value of a TSLS Estimator for a Single Mediator and Multiple Instruments

Proposition D.3: Given Assumptions D.1-D.4, the expected value of the TSLS estimator for the situation described in Equations D.1-D.3, $\hat{\beta}_{TSLS}^{(K)}$ is:

$$E\{\hat{\beta}_{TSLS}^{(K)}\} \approx \frac{TIV^{(K)}}{TIV^{(K)} + EIV^{(K)}}\beta_{ca} + \frac{EIV^{(K)}}{TIV^{(K)} + EIV^{(K)}}\beta_{cs} \qquad \text{(D.9)}$$

The same conclusion holds for the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k.$

Proof of this proposition is provided at the end of this appendix.

V. Mean and Median Bias in a TSLS Estimator for a Single Mediator with Multiple Instruments

Recall that we have defined the finite sample bias for a TSLS estimator as the difference between its expected value and β_{ca} (Equation 32a). Using this definition:

$$BIAS_{TSLS}^{(K)} \equiv E\{\hat{\beta}_{TSLS}^{(K)}\} - \beta_{ca} \approx \frac{EIV^{(K)}}{TIV^{(K)} + EIV^{(K)}}[\beta_{cs} - \beta_{ca}] \qquad \text{(D.10)}$$

As shown in Appendix B, the difference between β_{cs} and β_{ca} is the same as "OLS bias." Using this definition, the relationship between finite sample bias for a TSLS estimator with K instruments and the OLS bias can be expressed as:

$$BIAS_{TSLS}^{(K)} \approx \frac{EIV^{(K)}}{TIV^{(K)} + EIV^{(K)}}BIAS_{OLS} \qquad \text{(D.11)}$$

This result holds for the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k.$

Note that an overidentified TSLS estimator with K instruments has an asymptotically normal sampling distribution (Angrist and Pischke, 2009, p. 140). Thus its asymptotic mean equals its asymptotic median, and we can rewrite Equation D.9 as:

$$MEDIAN\{\hat{\beta}_{TSLS}^{(K)}\} \approx \frac{TIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}\beta_{ca} + \frac{EIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}\beta_{cs} \qquad \text{(D.12)}$$

Further note that we have defined the median bias of a TSLS estimator as the difference between its median and the true causal coefficient (Equation 32b). Using this definition and Equation D.9, we have:

$$\begin{aligned} MEDIANBIAS_{TSLS}^{(K)} \equiv MEDIAN\{\hat{\beta}_{TSLS}^{(K)}\} - \beta_{ca} &\approx \frac{EIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}[\beta_{cs}-\beta_{ca}] \\ &\approx \frac{EIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}BIAS_{OLS} \qquad \text{(D.13)} \end{aligned}$$

VI. When Are Multiple Instruments Worthwhile?

To answer this question, we first derive the expression for the population F-value for the first-stage regression with K instruments and use the F-value to measure the “strength” of the K instruments.

Proposition D.4: Based on Proposition B.2, it can be shown that the F-value for a first-stage regression with K instruments is:

$$\begin{aligned} F_{pop}^{(K)} &= \frac{E\left(tiv^{(K)}\right)+E\left(eiv^{(K)}\right)}{E\left(eiv^{(K)}\right)} \\ &= \frac{\left[N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+K\sigma_{M*}^2\right]/K}{\sigma_{M*}^2} \qquad \text{(D.14)} \end{aligned}$$

Proof for this proposition is provided at the end of the appendix.

Note that for the special case where $\phi = 0, and\ \pi_k = \pi\ for\ all\ k,$

$$\begin{aligned} F_{pop}^{(K)} &= \frac{\left[N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+K\sigma_{M*}^2\right]/K}{\sigma_{M*}^2} \\ &= \frac{\left[N\bar{T}(1-\bar{T})(\pi^2)+K\sigma_{M*}^2\right]/K}{\sigma_{M*}^2} \qquad \text{(D.15)} \end{aligned}$$

Combining Equation D.14 with Equations D.11 and D.13 yields:

$$BIAS_{TSLS}^{(K)} = MEDIANBIAS_{TSLS}^{(K)} \approx \frac{1}{F_{pop}^{(K)}}BIAS_{OLS} \qquad \text{(D.16)}$$

Using K instruments is worthwhile if they reduce the magnitude of the median TSLS bias when a single instrument is used, that is, if

$$MEDIANBIAS_{TSLS}^{(K)} < MEDIANBIAS_{TSLS}^{(1)}$$

which is equivalent to

$$F_{pop}^{(K)} > F_{pop}^{(1)}$$

where $F_{pop}^{(1)}$ is the population F-value for the single instrument case shown in Appendix B (Equation B.26).

Proposition D.5: For K instruments to be worthwhile, the first-stage F-statistics for K instruments needs to exceed the first-stage F-statistics for one instrument, which is equivalent to:

$$VAR(\pi_k) > (K-1)E(\pi_k)^2 \qquad \text{(D.17)}$$

For the special case of $\phi = 0, and\ \pi_k = \pi\ for\ all\ k$, this condition will not be met, i.e., when $\phi = 0, and\ \pi_k = \pi$

$$VAR(\pi) = 0 \leq (K-1)E(\pi)^2 \qquad \text{(D.18)}$$

In other words, for the special case where the true treatment effect is constant across strata, the magnitude of the TSLS bias for K instruments is larger than that for a single instrument.

Proof for this proposition is provided at the end of the appendix.

Proofs of Propositions D.1-D.5

Proof of Proposition D.1

By definition,

$$E(\pi_k) = \frac{1}{K}\pi_1 + \frac{1}{K}\pi_2 + \cdots + \frac{1}{K}\pi_K$$

$$= \frac{1}{K}\sum_{k=1}^{K}\pi_k$$

$$= \frac{1}{K}\sum_{k=1}^{K}[\pi_1 + (k-1)\phi]$$

$$= \frac{1}{K} * K * \pi_1 + \frac{1}{K}\sum_{k=1}^{K}(k-1)\phi$$

$$= \pi_1 + \frac{1}{K}\sum_{k=1}^{K}(k-1)\phi$$

$$= \pi_1 + \frac{1}{K} * \phi * [0 + 1 + 2 + \cdots + (K-1)]$$

It is well known that the sum of an integer series n is the following:[47]

$$\sum_{n=1}^{N} n = \frac{N(N+1)}{2}$$

Therefore, the above equation yields:

$$E(\pi_k) = \pi_1 + \frac{1}{K} * \phi * \frac{K(K-1)}{2}$$

$$= \pi_1 + \frac{K-1}{2}\phi \qquad \text{(A-D.1)}$$

Also by definition,

$$VAR(\pi_k) = E\{[\pi_k - E(\pi_k)]^2\}$$

$$= E\{\left[\pi_1 + (k-1)\phi - (\pi_1 + \frac{K-1}{2}\phi)\right]^2\}$$

$$= E\{\left[(k-1)\phi - (\frac{K-1}{2}\phi)\right]^2\}$$

[47] This result can be found discussed in the Web site: http://www.wikihow.com/Sum-the-Integers-from-1-to-N.

$$= E\{\left[\frac{2k-K-1}{2}\phi\right]^2\}$$

$$= \frac{\phi^2}{4}E\{[2k-K-1]^2\}$$

$$= \frac{\phi^2}{4}E\{4k^2-4kK-4k+K^2+2K+1\}$$

$$= \frac{\phi^2}{4}E\{4(k^2-kK-k)+(K+1)^2\}$$

$$= \phi^2E[(k^2-kK-k)]+\frac{\phi^2}{4}[E(K+1)^2]$$

$$= \phi^2E(k^2)-\ \phi^2(K+1)E(k)+\frac{\phi^2}{4}(K+1)^2$$

$$= \phi^2*\frac{1}{K}\sum_{k=1}^{K}k^2-\phi^2(K+1)*\frac{1}{K}\sum_{k=1}^{K}k+\frac{\phi^2}{4}(K+1)^2 \qquad \text{(A-D.2)}$$

Using the identity N^2 = N(N+1)/2 + (N-1)N/2 and mathematical induction, it can be shown that the sum of the first *k* square numbers is equal to $\frac{K(K+1)(2K+1)}{6}$.[48] Therefore, it follows that:

$$VAR(\pi_k) = \phi^2*\frac{1}{K}\frac{K(K+1)(2K+1)}{6}-\phi^2(K+1)*\frac{1}{K}\frac{K(K+1)}{2}+\frac{\phi^2}{4}(K+1)^2$$

$$= \phi^2*\frac{(K+1)(2K+1)}{6}-\frac{\phi^2}{2}(K+1)^2+\frac{\phi^2}{4}(K+1)^2$$

$$= \phi^2[\frac{(K+1)(2K+1)}{6}-\frac{(K+1)^2}{4}]$$

$$= \phi^2[(K+1)(\frac{(2K+1)}{6}-\frac{(K+1)}{4})]$$

$$= \phi^2[(K+1)(\frac{2*(2K+1)}{12}-\frac{3*(K+1)}{12})]$$

$$= \phi^2[(K+1)\left(\frac{K-1}{12}\right)]$$

$$= \phi^2(\frac{K^2-1}{12}) \qquad \text{(A-D.3)}$$

Therefore, the mean and variance of π_k are:

[48] The first verbal proof of this identity was credited to *Introduction to Arithmetic by* Nicomachus of Gerasa (c 100 A.D.). A mathematical proof of it can be found at http://pirate.shu.edu/~wachsmut/ira/infinity/answers/sm_sq_cb.html.

$$E(\pi_k) = \pi_1 + \frac{K-1}{2}\phi$$

$$VAR(\pi_k) = \phi^2(\frac{K^2-1}{12})$$

Proof of Proposition D.2

$$E\left(tiv^{(K)}\right) = N\bar{T}(1-\bar{T})\frac{1}{K}[\textstyle\sum_{k=1}^{K}(\pi_k)^2]$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}({\pi_1}^2 + {\pi_2}^2 + {\pi_3}^2 + \cdots + {\pi_K}^2)$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{{\pi_1}^2 + (\pi_1+\phi)^2 + (\pi_1+2\phi)^2 + \cdots + [\pi_1 + (K-1)\phi]^2\}$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{K{\pi_1}^2 + [2\pi_1\phi + 4\pi_1\phi + \cdots + 2(K-1)\pi_1\phi] + \phi^2[1^2 + 2^2 + \cdots + (K-1)^2]\}$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{K{\pi_1}^2 + 2\pi_1\phi[1 + 2 + \cdots + (K-1)] + \phi^2[1^2 + 2^2 + \cdots + (K-1)^2]\}$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{K{\pi_1}^2 + 2\pi_1\phi[\textstyle\sum_{k=1}^{K-1}k] + \phi^2[\sum_{k=1}^{K-1}k^2]\}$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{K{\pi_1}^2 + 2\pi_1\phi\frac{(K-1)(1+K-1)}{2} + \phi^2\frac{(K-1)(K-1+1)[2(K-1)+1]}{6}\}$$

$$= N\bar{T}(1-\bar{T})\frac{1}{K}\{K{\pi_1}^2 + (K-1)K\pi_1\phi + \phi^2\frac{K(K-1)(2K-1)}{6}\}$$

$$= N\bar{T}(1-\bar{T})\{{\pi_1}^2 + (K-1)\pi_1\phi + \phi^2\frac{(K-1)(2K-1)}{6}\} \qquad \text{(A-D.4)}$$

Proof of Proposition D.3

Consider the second-stage regression given in Equation D.2:

$$Y_{ik} = \alpha_k + \beta M_{ik} + \upsilon_{ik} \qquad \text{(D.2 restated)}$$

In this model, α_k represents the K strata-specific intercepts or strata fixed effects. One way of estimating the model is using a "fixed effects transformation" or "within transformation" to remove these strata fixed effects. (Wooldridge, 2002). Also referred to as "demeaning," this process entails averaging Equation D.2 over all units in a stratum to find stratum-level mean values of all terms in the model and subtracting the resulting model from the one in Equation D.2. Specifically, averaging Equation D.2 over all units in the stratum yields:

$$\overline{Y}_k = \alpha_k + \beta\overline{M}_k + \overline{\upsilon}_k \qquad \text{(A-D.5)}$$

where $\overline{Y}_k$, $\overline{M}_k$, and $\overline{\upsilon}_k$ represent stratum-level mean values of the outcome, mediator, and the unit-level error term, respectively. Subtracting Equation A-D.5 from Equation D.2 provides the demeaned model as:

$$Y_{ik} - \overline{Y}_k = \beta(M_{ik} - \overline{M}_k) + (\upsilon_{ik} - \overline{\upsilon}_k) \qquad \text{(A-D.6)}$$

or using Wooldridge's (2002) notation:

$$\ddot{Y}_{ik} = \beta\ddot{M}_{ik} + \ddot{\upsilon}_{ik} \qquad \text{(A-D.7)}$$

where $\ddot{Y}_{ik}$, $\ddot{M}_{ik}$, and $\ddot{v}_{ik}$ represent demeaned values of the outcome, mediator, and the unit-level error term, respectively.

As Wooldridge (2002) shows, estimation of Equation A-D.7 by OLS provides a consistent estimate of β. Hence, replacing the demeaned value of the mediator in Equation A-D.7 by the demeaned predicted values of the mediator ($\ddot{\hat{M}}_{ik} \equiv \hat{M}_{ik} - \overline{\hat{M}}_k$) and estimating the resulting regression by OLS yields the TSLS estimate of the causal coefficient with K instruments (as $\hat{M}_{ik}$ is based on the use of K instruments in the first stage). More formally:

$$\begin{aligned}\hat{\beta}^{(K)}_{TSLS} &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\ddot{Y}_{ik} - \overline{\ddot{Y}}_{ik})(\ddot{\hat{M}}_{ik} - \overline{\ddot{\hat{M}}}_{ik})}{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\ddot{\hat{M}}_{ik} - \overline{\ddot{\hat{M}}}_{ik})} \\ &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{N/K}\ddot{Y}_{ik}\ddot{\hat{M}}_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{N/K}\ddot{\hat{M}}_{ik}{}^{2}} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(Y_{ik} - \overline{Y}_k)(\hat{M}_{ik} - \overline{\hat{M}}_k)}{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\hat{M}_{ik} - \overline{\hat{M}}_k)^2}\end{aligned} \qquad \text{(A-D.8)}$$

where $\overline{\ddot{Y}}_{ik}$ and $\overline{\ddot{\hat{M}}}_k$ represent sample mean values of the demeaned outcome and predicted mediator, which are zero by construction. Next, following Hahn and Hausman (2002), we approximate the expected value of $\hat{\beta}^{(K)}_{TSLS}$ as the ratio of the expected value of the numerator in Equation A-D.8 to that of the denominator, as in Appendix B:

$$E\{\hat{\beta}^{(K)}_{TSLS}\} \approx \frac{E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(Y_{ik}-\overline{Y}_k)(\hat{M}_{ik}-\overline{\hat{M}}_k)\right\}}{E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}}_k)^2\right\}} \tag{A-D.9}$$

Let’s consider the numerator and denominator of the expression in Equation A-D.9 separately.

Numerator of the estimator

Note that we can rewrite the numerator of the expression in Equation A-D.9 as:

$$E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(Y_{ik}-\overline{Y}_k)(\hat{M}_{ik}-\overline{\hat{M}}_k)\right\} = \sum_{k=1}^{K}E\left\{\sum_{i=1}^{N/K}(Y_{ik}-\overline{Y}_k)(\hat{M}_{ik}-\overline{\hat{M}}_k)\right\} \tag{A-D.10}$$

It is important to recognize that Equation A-D.10 implies (i) the evaluation of the expression in the expectation separately within each stratum and (ii) summation of the results over K strata. This is consistent with considering each stratum as a separate randomized trial and pooling findings across them. Hence, given Assumptions D.1-D.4, the expression for the expected value of the numerator of the TSLS estimator derived in Appendix B for a sample of N units without any stratification applies to each stratum considered here. Recognizing that each stratum has N/K units and a treatment effect on the mediator of π_k, adapting Equation A-B.14 accordingly and substituting the result in Equation A-D.10 yields:

$$\begin{aligned}
&\sum_{k=1}^{K}E\left\{\sum_{i=1}^{N/K}(Y_{ik}-\overline{Y}_k)(\hat{M}_{ik}-\overline{\hat{M}}_k)\right\} = \sum_{k=1}^{K}\left[[\frac{N}{K}\overline{T}(1-\overline{T})\pi_k^2]\beta_{ca}+\sigma^2_{M_*}\beta_{cs}\right] \\
&= \frac{N}{K}\overline{T}(1-\overline{T})\beta_{ca}\sum_{k=1}^{K}\pi_k^2 + K\sigma^2_{M_*}\beta_{cs} \\
&= \frac{N}{K}\overline{T}(1-\overline{T})\beta_{ca}\left[K\pi_1^2+K(K-1)\pi_1\phi+\phi^2\frac{K(K-1)(2K-1)}{6}\right]+K\sigma^2_{M_*}\beta_{cs} \qquad \text{(A-D.11)} \\
&= N\overline{T}(1-\overline{T})\beta_{ca}\left[\pi_1^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\right]+K\sigma^2_{M_*}\beta_{cs}
\end{aligned}$$

where in the third line, we used the expression derived for $\sum_{k=1}^{K}\pi_k^2$ in the proof of Proposition D.2.

Substituting Equation A-D.11 in Equation A-D.10 yields:

$$E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(Y_{ik}-\overline{Y}_k)(\hat{M}_{ik}-\overline{\hat{M}}_k)\right\} = N\overline{T}(1-\overline{T})\beta_{ca}\left[\pi_1^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\right]+K\sigma^2_{M_*}\beta_{cs} \tag{A-D.12}$$

Denominator of the estimator

We can rewrite the denominator of the expression in Equation A-D.9 as:

$$E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}_{k}})^{2}\right\}=\sum_{k=1}^{K}E\left\{\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}_{k}})^{2}\right\} \qquad \text{(A-D.13)}$$

As for the numerator, the expression derived in Appendix B for the expected value of the denominator of the TSLS estimator applies for the expression $E\left\{\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}_{k}})^{2}\right\}$, which is essentially the expected value of the sum of squares of the predicted mediator values for each stratum. Hence, adapting Equation A-B.17 for a sample of NIK units and substituting it in Equation A-D.13 yields:

$$\begin{aligned}
E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}_{k}})^{2}\right\}&=\sum_{k=1}^{K}E\left\{\sum_{i=1}^{N/K}(\hat{M}_{ik}-\overline{\hat{M}_{k}})^{2}\right\}\\
&=\sum_{k=1}^{K}\left[\frac{N}{K}\overline{T}(1-\overline{T})\pi_{k}^{2}+\sigma_{M*}^{2}\right]\\
&=\frac{N}{K}\overline{T}(1-\overline{T})\sum_{\text{k}=1}^{K}\pi_{k}^{2}+K\sigma_{M*}^{2} \qquad \text{(A-D.14)}\\
&=\frac{N}{K}\overline{T}(1-\overline{T})\left[K\pi_{1}^{2}+K(K-1)\pi_{1}\phi+\phi^{2}\,\frac{K(K-1)(2K-1)}{6}\right]+K\sigma_{M*}^{2}\\
&=N\overline{T}(1-\overline{T})\left[\pi_{1}^{2}+(K-1)\pi_{1}\phi+\phi^{2}\,\frac{(K-1)(2K-1)}{6}\right]+K\sigma_{M*}^{2}
\end{aligned}$$

Full expression

Substituting Equations A-D.12 and A-D.14 in Equation A-D.9 yields:

$$E\{\hat{\beta}_{TSLS}^{(K)}\}\approx\frac{N\overline{T}(1-\overline{T})\beta_{ca}\left[\pi_{1}^{2}+(K-1)\pi_{1}\phi+\phi^{2}\,\frac{(K-1)(2K-1)}{6}\right]+K\sigma_{M*}^{2}\beta_{cs}}{N\overline{T}(1-\overline{T})\left[\pi_{1}^{2}+(K-1)\pi_{1}\phi+\phi^{2}\,\frac{(K-1)(2K-1)}{6}\right]+K\sigma_{M*}^{2}} \qquad \text{(A-D.15)}$$

Substituting Equations D.7 and D.8 in Equation A-D.15 yields:

$$E\{\hat{\beta}_{TSLS}^{(K)}\}\approx\frac{TIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}\beta_{ca}+\frac{EIV^{(K)}}{TIV^{(K)}+EIV^{(K)}}\beta_{cs} \qquad \text{(A-D.16)}$$

For the special case where $\phi=0$ and $\pi_{\text{k}}=\pi$, Equation A-D.16 becomes:

$$E\{\hat{\beta}_{TSLS}^{(K)}\} \approx \frac{N\overline{T}(1-\overline{T})\beta_{ca}\pi + K\sigma_{M*}^{2}\beta_{cs}}{N\overline{T}(1-\overline{T})\pi_{1}^{2} + K\sigma_{M*}^{2}} \qquad \text{(A-D.17)}$$

Adapting Equations D.7 and D.8 for this special case and substituting them in Equation A-D.17 yields the same expression in A-D.16.

Proof of Proposition D.4

In the situation under consideration in this appendix, there is a randomized trial for each of K strata, and by interacting treatment status with a dichotomous indicator for each stratum and pooling data across strata, one can create K instrumental variables to be used in the following first-stage regression:

$$M_{ik} = \sum_{m=1}^{K} \mu_m S_k^{(m)} + \sum_{m=1}^{K} \pi_m S_k^{(m)} T_{ik} + \varepsilon_{ik} \qquad \text{(A-D.18)}$$

where $S_{ik}^{(m)}$equals one when m equals k and zero otherwise.

To simplify the proof, it is assumed that each stratum is independent of the others (its units are sampled, randomized, and treated separately). In addition, we assume that all strata have the same total number of units, $(N_k = N/K)$, proportion of units randomized to treatment, $\overline{T}_k$, and variance of counter-factual mediator values, $\sigma_{M*(K)}^{2}$.

Therefore, the above regression can be viewed as a combination of K independent regressions like the following, one for each stratum:

$$M_{ik} = \mu_k + \pi_k T_{ik} + \varepsilon_{ik} \qquad \text{(A-D.19)}$$

Recalling Equation A-B.24 from Appendix B and adapting it for the regression under consideration shows that the first-stage F-value is approximately the following:

$$F_{pop}^{(K)} = E\left(F_{sample}\right) = E\left(\frac{SS(\widehat{M_{\mathrm{ik}}})/(2K-K)}{SS(\widehat{\varepsilon_{\mathrm{ik}}}))/(N-2K)}\right) \approx \frac{E(SS(\widehat{M_{\mathrm{ik}}}))/(K)}{E(SS(\widehat{\varepsilon_{\mathrm{ik}}}))/(N-2K)} \qquad \text{(A-B.24 restated)}$$

For the first-stage regression represented by Equation A-D.18, the expected value of the sum of squares predicted by the regression is then essentially the sum (across K strata) of the expected values of the sum of squares for the predicted value of the mediator for each stratum. That is,

$$E[SS(\widehat{M}_{ik})] = \sum_{k=1}^{K} E\{\sum_{i=1}^{n} (\widehat{M}_{ik} - \overline{\widehat{M}}_k)^2\} \qquad \text{(A-D.20)}$$

Recall from Equation A-B.22 that, if there is no strata within the full sample, then:

$$E(SS_p) = E(SS(\widehat{M_{\mathrm{i}}})) = E\{\sum_{i=1}^{N}\left(\widehat{M_{\mathrm{i}}} - \overline{\widehat{M}}\right)^2\}$$

$$= \sigma_{M*}^{2} + N\overline{T}(1-\overline{T})\pi^{2} \qquad \text{(A-D.21)}$$

Therefore, for a given stratum k,

$$E(SS(\widehat{M_{\imath k}})) = \sigma^2_{M*} + \frac{N}{K}\bar{T}(1-\bar{T})\pi_k^2 \text{ for a given k} \qquad \text{(A-D.22)}$$

Substituting Equation A-D.22 into Equation A-D.20 yields:

$$E[SS(\widehat{M}_{ik})] = \sum_{k=1}^{K}[N_k\bar{T}(1-\bar{T})\pi_k^2 + \sigma^2_{M*}]$$

$$= \sum_{k=1}^{K}[\frac{N}{K}\bar{T}(1-\bar{T})\pi_k^2 + \sigma^2_{M*}]$$

$$= \frac{N}{K}\bar{T}(1-\bar{T})\sum_{k=1}^{K}[\pi_k^2] + K\sigma^2_{M*} \qquad \text{(A-D.23)}$$

Recall from Section III that:

$E\left(tiv^{(K)}\right) = N\bar{T}(1-\bar{T})\frac{1}{K}[\sum_{k=1}^{K}(\pi_k)^2]$, and

$E\left(eiv^{(K)}\right) = K\sigma^2_{M*}$

It follows that:

$$E[SS(\widehat{M}_{ik})] = E\left(tiv^{(K)}\right) + E\left(eiv^{(K)}\right) \qquad \text{(A-D.24)}$$

Similarly, based on the assumption that the variance of counterfactual mediator values, $\sigma^2_{M*(K)}$is the same across all strata, it is denoted by σ^2_{M*} for all strata. It then follows that:

$$E\left(SS(\hat{\varepsilon}_{ik})\right) = KE[\sum_{i=1}^{N/K}(M_{ik} - \widehat{M}_{ik})^2]$$

$$= KE[\sum_{i=1}^{N/K}(M_{ik} - \widehat{M}_{ik})^2]$$

$$= K * \left(\frac{N}{K} - 2\right) VAR(\hat{\varepsilon}_{ik})$$

$$= K * \left(\frac{N}{K} - 2\right)\sigma^2_{M*}$$

$$= \left(\frac{N}{K} - 2\right)E\left(eiv^{(K)}\right) \qquad \text{(A-D.25)}$$

Substituting Equations A-D.24 and A-D.25 into Equation A-B.24 yields:

$$F_{pop}^{(K)} \approx \frac{E(SS(\widehat{M_{\mathrm{ik}}}))/K}{E(SS(\widehat{\varepsilon_{\mathrm{ik}}}))/(N-2K)}$$

$$= \frac{[E\left(tiv^{(K)}\right) + E\left(eiv^{(K)}\right)]/K}{[\left(\frac{N}{K} - 2\right)E\left(eiv^{(K)}\right)]/(N-2K)}$$

$$= \frac{[E(tiv^{(K)})+E(eiv^{(K)})]/K}{[(\frac{N-2K}{K})E(eiv^{(K)})]/(N-2K)}$$

$$= \frac{[E(tiv^{(K)})+E(eiv^{(K)})]}{E(eiv^{(K)})} \tag{A-D.26}$$

Substituting Equations D.7 and D.8 into Equation A-D.26 yields:

$$F_{pop}^{(K)} = \frac{E(tiv^{(K)})+E(eiv^{(K)})}{E(eiv^{(K)})}$$

$$= \frac{\left[N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+K\sigma_{M_*}^2\right]/K}{\sigma_{M_*}^2} \tag{A-D.27}$$

Proof of Proposition D.5

For K instruments to be worthwhile, the first-stage F-statistic for K instruments needs to exceed the first-stage F-statistics for one instrument, that is,

$$F_{pop}^{(K)} > F_{pop}^{(1)}$$

Substituting Equation D.12 into the above inequality yields:

$$\frac{\left[N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+K\sigma_{M_*}^2\right]/K}{\sigma_{M_*}^2} > \frac{N\bar{T}(1-\bar{T})(\pi_1+\frac{K-1}{2}\phi)^2+\sigma_{M_*}^2}{\sigma_{M_*}^2}$$

Since $\sigma_{M_*}^2 > 0$ and $\bar{T}(1-\bar{T}) \geq 0$, this inequality can be further simplified to:

$$\left[N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+K\sigma_{M_*}^2\right]/K > N\bar{T}(1-\bar{T})(\pi_1+\frac{K-1}{2}\phi)^2+\sigma_{M_*}^2$$

→

$$\frac{1}{K}N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}+\sigma_{M_*}^2 > N\bar{T}(1-\bar{T})(\pi_1+\frac{K-1}{2}\phi)^2+\sigma_{M_*}^2$$

→

$$\frac{1}{K}N\bar{T}(1-\bar{T})\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}>N\bar{T}(1-\bar{T})(\pi_1+\frac{K-1}{2}\phi)^2$$

→

$$\{{\pi_1}^2+(K-1)\pi_1\phi+\phi^2\frac{(K-1)(2K-1)}{6}\}>K(\pi_1+\frac{K-1}{2}\phi)^2$$

→

$$\{(\pi_1+\frac{K-1}{2}\phi)^2+\phi^2(\frac{K^2-1}{12})\}>K(\pi_1+\frac{K-1}{2}\phi)^2$$

Recall from Equations D.4 and D.5 that:

$$E(\pi_k)=\ \pi_1+\tfrac{K-1}{2}\phi \qquad \text{(D.4 restated)}$$

$$VAR(\pi_k)=\ \phi^2(\tfrac{K^2-1}{12}) \qquad \text{(D.5 restated)}$$

Substituting Equations D.4 and D.5 into the above inequality yields:

$$\left\{(\pi_1+\frac{K-1}{2}\phi)^2+\phi^2\left(\frac{K^2-1}{12}\right)\right\}=\{E(\pi_k)^2+VAR(\pi_k)\}>K(\pi_1+\frac{K-1}{2}\phi)^2=KE(\pi_k)^2$$

→

$$E(\pi_k)^2+VAR(\pi_k)>\ KE(\pi_k)^2$$

→

$$VAR(\pi_k)>(K-1)E(\pi_k)^2$$

References

Anderson, T. W., and H. Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20: 46-63.

Angrist, J., and A. Krueger. 1991. "Does Compulsory Schooling Attendance Affect Schooling and Earnings." *Quarterly Journal of Economics* 106: 979-1014.

Angrist, J., and A. Krueger. 1999. "Empirical Strategies in Labor Economics." In Orley Ashenfelter and David Card (eds.), *The Handbook of Labor Economics.* Amsterdam, The Netherlands: Elsevier Science B. V.

Angrist, J., and S. Pischke. 2008. *Mostly Harmless Econometrics*, 1st Edition. NJ: Princeton University Press.

Basman, R. L. 1960. "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics." *Journal of the American Statistical Association* 55: 650-659.

Basman, R. L. 1963. "Remarks Concerning the Application of Exact Finite Sample Distribution Functions of GCL Estimators in Econometric Statistical Inference." *Journal of the American Statistical Association* 58: 943-976.

Blalock, Hubert. 1972. *Social Statistics*, 1st Edition. New York: McGraw-Hill.

Bloom, Howard S. 2005. "Randomizing Groups to Evaluate Place-Based Programs." In Howard S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bound, J., A. Jaeger, and R. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90: 443-450.

Brennan, Robert L. 2001. *Generalizability Theory*. New York: Springer.

Chamberlain, G., and G. Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica* 72, 1: 295-306.

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles.* New York: John Wiley.

Deaton, Angus. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, Maryland: The Johns Hopkins University Press.

Gamse, B. C., R. T. Jacob, M. Horst, B. Boulay, and F. Unlu. 2008. *Reading First Impact Study Final Report* (NCEE 2009-4038).Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Greene, William. 1997. *Econometric Analysis*, 3rd Edition. NJ: Prentice Hall.

Guilford, J. P. 1965. *Fundamental Statistics in Psychology and Education,* 4th Edition. NY: McGraw-Hill.

Hahn, Jinyong, and Jerry Hausman. 2002. “A New Specification Test for the Validity of Instrumental Variables.” *Econometrica* 70:1: 163-189.

Institute of Education Sciences, U.S. Department of Education. 2008. *Rigor and Relevance Redux: Director’s Biennial Report to Congress* (IES 2009-6010). Washington, DC: U.S. Government Printing Office.

Jackson, Russell, Ann McCoy, Carol Pistorino, Anna Wilkinson, John Burghardt, Melissa Clark, Christine Ross, Peter Schochet, and Paul Swank. 2007. *National Evaluation of Early Reading First: Final Report.* U.S. Department of Education, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office.

Kling, Jeffrey, Jeffrey B. Liebman, and Lawrence F. Katz. 2007. “Experimental Analysis of Neighborhood Effects.” *Econometrica* 75, 1: 83-119.

Ludwig, Jens, and Jeffrey R. Kling. 2007. “Is Crime Contagious?” *Journal of Law and Economics* 50, 3: 491-518.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research,* 1st Edition. NY: Cambridge University Press.

Nelson, Charles R., and Richard Startz. 1990. “The Distribution of the Instrumental Variables Estimator and Its T-Ratio When the Instrument Is a Poor One.” *Journal of Business* 63 (1): 125-140.

Pianta, R. C., K. M. La Paro, C. Payne, M. Cox, and R. Bradley. 2002. “The Relation of Kindergarten Classroom Environment to Teacher, Family, and School Characteristics and Child Outcomes.” *Elementary School Journal* 102, 3: 225-238.

Raudenbush, Stephen W. 2007. “Statistical Inference When Classroom Quality Is Measured with Error.” Paper prepared for a meeting on “Approaches to Assessing Classroom Quality,” jointly sponsored by the Spencer Foundation and the William T. Grant Foundation on February 21, 2007, in Chicago.

Raudenbush, Stephen W., Andres Martinez, Howard S. Bloom, Pei Zhu, and Fen Lin. 2008. “The Reliability of Group-Level Measures and the Power of Group-Randomized Studies.” Working Paper. New York: MDRC.

Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66, 5: 688-701.

Rubin, Donald B. 1977. “Assignment to a Treatment Group on the Basis of a Covariate.” *Journal of Education Statistics* 2: 1-26.

Rubin, Donald B. 1978. “Bayesian Inference for Causal Effects: The Role of Randomization.” *The Annals of Statistics* 6: 34-58.

Shavelson, Richard J., and Noreen M. Webb. 1991. *Generalizability Theory: A Primer.* Newbury Park, CA: Sage Publications.

Stock, J. H., and M. Yogo. 2005. “Testing for Weak Instruments in Linear IV Regression.” In D. W. K. Andrews and J. H. Stock (eds.), *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg.* Cambridge, UK: Cambridge University Press.

Wald, A. 1940. “The Fitting of Straight Lines if Both Variables Are Subject to Error.” *Annals of Mathematical Statistics* 11, 284–300.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.