Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Bias and Bias Correction in Multisite Instrumental Variables Analysis of Heterogeneous Mediator Effects

Sean F. Reardon, Fatih Unlu, Pei Zhu and Howard S. Bloom

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2014 39: 53 originally published online 16 December 2013
DOI: 10.3102/1076998613512525

The online version of this article can be found at:
http://jeb.sagepub.com/content/39/1/53

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
2014, Vol. 39, No. 1, pp. 53–86
DOI: 10.3102/1076998613512525

Bias and Bias Correction in Multisite Instrumental Variables Analysis of Heterogeneous Mediator Effects

Sean F. Reardon
Stanford University

Fatih Unlu
Abt Associates

Pei Zhu
Howard S. Bloom
MDRC

We explore the use of instrumental variables (IV) analysis with a multisite randomized trial to estimate the effect of a mediating variable on an outcome in cases where it can be assumed that the observed mediator is the only mechanism linking treatment assignment to outcomes, an assumption known in the IV literature as the exclusion restriction. We use a random-coefficient IV model that allows both the impact of program assignment on the mediator (compliance with assignment) and the impact of the mediator on the outcome (the mediator effect) to vary across sites and to covary with one another. This extension of conventional fixed-coefficient IV analysis illuminates a potential bias in IV analysis which Reardon and Raudenbush refer to as "compliance-effect covariance bias." We first derive an expression for this bias and then use simulations to investigate the sampling variance of the conventional fixed-coefficient two-stage least squares (2SLS) estimator in the presence of varying (and covarying) compliance and treatment effects. We next develop two alternate IV estimators that are less susceptible to compliance-effect covariance bias. We compare the bias, sampling variance, and root mean squared error of these "bias-corrected IV estimators" to those of 2SLS and ordinary least squares (OLS). We find that, when the first-stage F*-statistic exceeds 10 (a commonly used threshold for instrument strength), the bias-corrected estimators typically perform better than 2SLS or OLS. In the last part of the article, we use both the new estimators and 2SLS to reanalyze data from two large multisite studies.*

Keywords: *mediator effect; instrumental variables; multisite trials*

I. Introduction

The large number of randomized trials and regression discontinuity analyses that have been conducted during the past decade have produced internally valid

estimates of the causal effects of many different social and educational interventions on many different types of behaviors and outcomes for many different types of individuals. These findings provide a growing base of credible evidence about the effectiveness of specific interventions, which is beginning to play an important role in evidence-based policymaking and practice. However, because the theories behind many interventions are not well developed, and because many interventions have multiple components, it is generally more complicated to determine the mechanisms through which an intervention operates.

Understanding the mechanisms through which an intervention operates requires identifying a set of hypothesized mediators through which the intervention operates, estimating the effects of the intervention on these mediators, and then estimating the effects of the mediators on the outcomes of interest. Although randomized experiments provide a straightforward method of estimating the effect of an intervention on a mediator, they do not provide as straightforward a method of obtaining unbiased estimates of the effect of a mediator on an outcome. This is both because the mediators are not randomly assigned (which leads to selection bias) and because the values of the mediators are often measured with error (which leads to measurement error–induced attenuation bias, also known as "errors-in-variables"). Under certain conditions, however, instrumental variables (IV) methods can be used to obtain unbiased estimates of mediator effects in randomized experiments or regression discontinuity analyses.

The intuition of the IV method is as follows: A randomized trial or regression discontinuity analysis can provide an internally valid estimate of the effects of an assigned treatment (T) on an outcome (Y) and on a mediator (M). In situations like this, the assigned treatment is an "instrument" of exogenous change in both the mediator and the outcome. In the simplest case, if it can be assumed that the full effect of the treatment on the outcome is produced by the mediator (an assumption known as the "exclusion restriction"), the average effect of the mediator on the outcome $\left(\frac{\Delta Y}{\Delta M}\right)$ equals the ratio of the effect of the treatment on the outcome $\left(\frac{\Delta Y}{\Delta T}\right)$ to the effect of the treatment on the mediator $\left(\frac{\Delta M}{\Delta T}\right)$. Because the randomized experiment or regression discontinuity design provides unbiased estimates of the latter two effects, their ratio will be an (asymptotically) unbiased estimate of the effect of a unit change in the mediator on the outcome $\left(\frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{\Delta Y}{\Delta M}\right)$.

Consider, for example, the recent multisite impact evaluation of the federal Reading First (RF) Program (Gamse, Bloom, Kemple, & Jacob, 2008) on reading achievement in the early elementary school grades. RF's theory of change posits that the RF program would increase teachers' use of five dimensions of reading instruction (phonemic awareness, phonics, vocabulary, fluency, and comprehension; hereafter referred to as "RF instructional methods") and that this type of instruction improves students' reading achievement. Because instructional methods were not randomized in the RF study, we can use an IV analysis to test

the latter hypothesis, under the assumption that the only way that assignment to RF would affect student achievement was through its effect on the amount of time teachers spent using the desired instructional methods. The results of the RF impact study showed that, on average, RF increased the amount of time that teachers spent on RF instruction by 11.6 min per day $\left(\frac{\Delta M}{\Delta T} = 11.6\right)$ and increased student reading achievement by 4.29 scale score points $\left(\frac{\Delta Y}{\Delta T} = 4.29\right)$. If all of RF's effect on reading achievement is produced by its effect on the use of RF instructional methods, these findings imply that the effect of such instruction is 0.37 scale score points per additional instructional minute $\left(\frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{4.29}{11.6} = 0.37\right)$.

The RF study was a multisite trial in which schools in 18 sites (17 school districts and 1 statewide program) were assigned on the basis of a continuous rating score or by randomization to receive the RF program or not. In a multisite design, a more complex IV analysis is possible. Because the treatment is ignorably assigned in each site, site-specific instruments can be constructed by interacting treatment assignment with a zero/one indicator for each site. Such "multiple-site, multiple-instrument" IV analyses can have both advantages and disadvantages.

One potential advantage is an increase in precision that will occur if the effect of treatment assignment on the mediator varies substantially across sites. For example, if RF increased the use of RF instruction by 20 min per daily reading block in some sites and by 2 min per daily reading block in other sites, an analysis that uses a separate instrument for each site can leverage this variation to provide more precise estimates of the average mediator effect. A second potential advantage of using a separate instrument for each site is that doing so may make it possible to study how the mediator effect *varies* across sites, if the sample sizes within each site are sufficiently large to enable precise estimates within each site. A third potential advantage of using a separate instrument for each site is that this makes it possible to study the separate effects of multiple mediators of a given intervention, as was done by Kling, Liebman, and Katz (2007), Duncan, Morris, and Rodrigues (2011), and Nomi and Raudenbush (2012).

A potential disadvantage of using multiple site-by-treatment interactions as instruments is that, if the impacts of the treatment on the mediator do not vary significantly across sites, the use of multiple instruments may lead to substantially decreased precision and increased finite sample bias (FSB; Angrist & Pischke, 2009; Bound, Jaeger, & Baker, 1995; Hahn & Hausman, 2002; Stock & Yogo, 2005).

In this article, we investigate the magnitude of the bias of multiple-site, multiple-instrument IV estimators. We consider not only the role of FSB but also the role of a second type of bias, what Reardon and Raudenbush (in press) refer to as "compliance-effect covariance bias." This bias arises if the effect of the treatment on the mediator and the effect of the mediator on the outcome covary across sites (or persons, though in the present article we are concerned with between-site variation).[1] Reardon and Raudenbush derive expressions for the value of

compliance-effect covariance bias under two-stage least squares (2SLS) estimation of multiple-site, multiple-instrument IV models with infinite samples, but do not examine compliance-effect covariance bias in finite samples. In this article, we extend Reardon and Raudenbush's analysis by deriving an expression for compliance-effect covariance bias of 2SLS in finite samples. We then conduct a set of simulations that explore the sampling variance of 2SLS estimates in the presence of compliance-effect covariance. We find that compliance-effect covariance bias can be substantial, that it grows asymptotically with sample size (unlike FSB, which declines with sample size), and that conventional 2SLS standard errors substantially underestimate the true sampling variance of the estimates when the effects of the mediator are heterogeneous.

In the second half of the article, we develop two "bias-corrected IV estimators" that are designed to reduce bias caused by compliance-effect covariance across sites. We use simulations to compare the statistical properties of these new estimators to those of 2SLS and ordinary least squares (OLS). These findings indicate that under a wide range of conditions, the new estimators perform better than 2SLS and OLS (in terms of bias and root mean squared error [RMSE]) if the instruments used have a first-stage F-statistic greater than 10 (a commonly recommended threshold for defining sufficiently "strong" instruments; see Staiger & Stock, 1997; Stock & Yogo, 2005).

The article concludes with two examples of the application of the bias-corrected IV estimators. We first use them to estimate the effect of class size on student achievement, using data from the Tennessee class-size experiment, Project Student–Teacher Achievement Ratio (STAR). We then use them to reanalyze data from the RF Impact Study described previously, estimating the per unit effect of RF instructional methods on students' reading achievement. These two empirical examples provide a useful contrast of potential applications.

II. Bias in the 2SLS Estimator

Notation

Consider a multisite randomized trial, in which N subjects (indexed by i) are nested in a set of K sites (indexed by $s \in \{1, 2, \ldots, K\}$). Within each site, a random sample of $n = N/K$ subjects (which can be individuals, classrooms, or schools) are ignorably assigned to treatment condition $T \in \{0, 1\}$. Let $p \in (0, 1)$ denote the proportion of subjects in each site assigned to the treatment condition $T = 1$. Note that, for ease of exposition, we set n and p to be constant across sites.

In each site, treatment status is assumed to affect an outcome Y through a single mediator M. Both the person-specific effect of T on M (the person-specific "compliance," denoted Γ) and the person-specific effect of M on Y (the person-specific "effect," denoted Δ) may be heterogeneous across subjects.

Our goal is to estimate the average effect of M on Y in the population, denoted by $\delta = E[\Delta]$.

Throughout the article, we make several assumptions. First, we make a pair of "stable unit treatment value assumptions," or SUTVA, described by Rubin (1986; see also Angrist, Imbens, & Rubin, 1996; Reardon & Raudenbush, in press, for statements of the SUTVA assumptions in the IV case). This is required so that the causal estimands are well defined. We also assume that $cov_s(\Gamma, \Delta) = [cov(\Gamma, \Delta)|S = s] = 0$ (no within-site compliance-effect covariance). Although implicit in all IV models where the mediator is not binary, this assumption is not trivial in many cases—in particular, it may be violated if individuals have some knowledge of the likely impact that M will have on them, and can choose levels of M in response to T, as in the Roy (1951) model. However, this assumption is met unambiguously if both T and M are binary and we focus only on compliers (Reardon & Raudenbush, in press). We assume no within-site compliance-effect covariance in order to focus on a distinct type of bias that may arise in multisite IV analyses. To that end, we do not assume that the between-site compliance-effect covariance (denoted $cov(\gamma_s, \delta_s)$, where the average compliance in site s is denoted γ_s and the average effect of M on Y in site s is denoted δ_s) is zero; our focus in this article is on the bias generated by nonzero covariance.

Within a given site s, let the data-generating model be

$$
\begin{aligned}
M_i &= \Lambda_s + \gamma_s T_i + e_i, \quad e_i \sim N\left(0, \sigma^2\right) \\
Y_i &= \Theta_s + \delta_s M_i + u_i, \quad u_i \sim N\left(0, \omega^2\right) \\
\begin{pmatrix} e_i \\ u_i \end{pmatrix} &\sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\omega \\ \rho\sigma\omega & \omega^2 \end{pmatrix} \right],
\end{aligned}
$$

where ρ is the correlation between e and u. Across sites, the covariance matrix of the γ_s's and the δ_s's is

$$
\begin{pmatrix} \gamma_s \\ \delta_s \end{pmatrix} \sim \left[\begin{pmatrix} \gamma \\ \delta \end{pmatrix}, \begin{pmatrix} \tau_\gamma & \tau_{\gamma\delta} \\ \tau_{\gamma\delta} & \tau_\delta \end{pmatrix} \right]. \tag{1}
$$

Note that the intercepts Λ_s and Θ_s here are conceived of as fixed (rather than random) and may be correlated with one another and/or with γ_s and δ_s. These intercepts are irrelevant to the bias, however, so it is not necessary to specify their structure.

Estimation

We wish to estimate $\delta = E[\Delta]$. One approach would be to estimate $\delta_s = E[\Delta|S = s]$ in each site separately, using standard IV methods, and then to average the δ_s's across sites. There are several drawbacks to this approach, however. First, if the instrument is weak in some sites, the estimated δ_s in those sites may be substantially biased due to FSB, leading to bias in the estimated average

effect. Second, a precision-weighted average of the δ_s will weight sites with greater compliance (larger values of γ_s) more, leading to biased estimates of δ if $\tau_{\gamma\delta} \neq 0$ (see Raudenbush, Reardon, & Nomi, 2012).

A second approach would be to pool the data across sites and fit a just-identified site-fixed effects IV model, using only a single instrument (Raudenbush et al., 2012). If γ_s is heterogeneous, such a model will be inefficient because it will not make use of all the exogenous variation in the mediator M that is induced by the instrument.

A third approach is to pool the data and fit an overidentified IV model, using K site-by-treatment status interactions as instruments. As we noted above, such a model may be preferable to either of the two approaches above in some cases. Because these instruments may collectively account for much more variation than a single instrument, the overidentified model may be more efficient than the single instrument model. In addition, by pooling the data, bias due to weak instruments in individual sites may be avoided.[2] Moreover, unlike the two approaches above, which can only be used if there is a single mediator, the multiple site-by-treatment interaction IV model can be used to identify the effects of multiple mediators. Although we do not consider the multiple mediator case in this article, our approach here may be adapted to that case.

We implement this approach as follows: First, we construct K instruments as site-by-treatment status interactions. Denote these as $Z_i^s = D_i^s T_i$, where $D_i^s = 1$ if subject i is in site s and $D_i^s = 0$ otherwise. Now the first-stage model is

$$M_i = \Lambda_s + \sum_{s=1}^{K} \gamma_s Z_i^s + e_i, \quad e_i \sim N\left(0, \sigma^2\right). \tag{2a}$$

The second-stage equation is

$$Y_i = \Theta_s + \delta \widehat{M}_i + u_i, \quad u_i \sim N\left(0, \omega^2\right). \tag{2b}$$

Bias in OLS and 2SLS Estimation

Now let F denote the population F-statistic (the expected value of the F-statistic corresponding to the null hypothesis that $\gamma_s = 0\ \forall s$ in the first-stage equation). We show in online Appendix A1 that this will be equal to

$$F = \frac{np(1-p)}{\sigma^2}\left(\gamma^2 + \tau_\gamma\right) + 1. \tag{3}$$

Estimating δ via OLS will lead to bias if M_i is correlated with u_i in Equation 2b. In online Appendix A2, we show that the OLS bias (the bias in the estimate of δ obtained from fitting Equation 2b via OLS) will be

$$E\left[\hat{\delta}^{\mathrm{OLS}}\right] - \delta = \rho\frac{\omega}{\sigma}\left(\frac{n}{F+n-1}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2+\tau_\gamma}\left(\frac{F-1}{F+n-1}\right). \tag{4a}$$

Estimating δ via 2SLS will also result in bias. In particular, as we show in online Appendix A3, the 2SLS bias (the bias in the estimate of δ obtained from fitting Equations 2a and 2b via 2SLS) is approximately

$$E\left[\hat{\delta}^{2SLS}\right] - \delta \approx \rho\frac{\omega}{\sigma}\left(\frac{1}{F}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\frac{F-1}{F}\right). \quad (5a)$$

Note that both the OLS bias and the 2SLS bias have two components—one component that depends on the covariance of the errors (ρ) and one component that depends on the covariance between the γs and Δs ($\tau_{\gamma\delta}$). The first component can be thought of as bias that arises from *treatment selection on levels* (individuals' received value of M is correlated with their potential value of Y that we would observe if they were assigned $M = 0$); it gives rise to selection bias in OLS and FSB in IV estimators. The second component can be thought of as bias that arises from *compliance selection on site-average effects* (site-average compliance with the instrument is correlated with the site-average effect the mediator has on the outcome Y), as might be predicted by the Roy model (Borjas, 1987; Roy, 1951); it gives rise to what we refer to as compliance-effect covariance bias (Reardon & Raudenbush, in press). Equations 4a and 5a make clear that both OLS and 2SLS are biased in finite samples if either $\rho \neq 0$ or $\tau_{\gamma\delta} \neq 0$. Moreover, both the OLS and the 2SLS biases can be written as weighted averages of the two components:

$$E\left[\hat{\delta}^{OLS}\right] - \delta = \rho\frac{\omega}{\sigma}\left(1 - \lambda^{OLS}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\lambda^{OLS}\right), \quad (4b)$$

and

$$E\left[\hat{\delta}^{2SLS}\right] - \delta \approx \rho\frac{\omega}{\sigma}\left(1 - \lambda^{2SLS}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\lambda^{2SLS}\right), \quad (5b)$$

where $\lambda^{OLS} = \frac{F-1}{F+n-1}$ and $\lambda^{2SLS} = \frac{F-1}{F}$. In the case of OLS, the weighting depends on the relative magnitudes of F and n. If $n \gg F$, λ^{OLS} approaches 0, in which case the bias due to the correlation of the errors is most significant. In the case of 2SLS, however, the weight depends only on the magnitude of F. When F is large, bias due to the correlation of the errors (FSB) is minimized and bias due to the correlation of γ_s and δ_s plays a dominant role. Because $\lambda^{2SLS} > \lambda^{OLS}$ for $n > 1$, the bias due to the second component will always get more weight in the 2SLS estimator than in the OLS estimator. However, the total bias will depend not just on these weights but on the relative magnitude of the two bias components. Thus, it is not a priori clear whether 2SLS yields less bias than OLS.

Factors Contributing to Bias in the 2SLS Estimator

The first component of bias in Equation 5a is pure FSB. This bias term is proportional to the within-site correlation of the error terms in the first- and

second-stage equations and inversely proportional to F. As F gets large, FSB becomes trivial.

The second component of the bias in Equation 5a is compliance-effect covariance bias. If $\gamma = 0$, this bias term is 0.[3] If, however, $\gamma \neq 0$, we can write the compliance-effect covariance bias term as

$$\frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\frac{F-1}{F}\right) = 2\text{Corr}(\gamma_s, \delta_s)\sqrt{\tau_\delta}\left(\frac{\text{CV}_\gamma}{\text{CV}_\gamma^2 + 1}\right) \times \left(\frac{F-1}{F}\right), \tag{6}$$

where $\text{CV}_\gamma = \sqrt{\tau_\gamma}/\gamma$ is the coefficient of variation of γ_s.

The compliance-effect covariance bias component depends on four factors. First, the bias term is proportional to the correlation between γ_s and δ_s. Second, the bias term is proportional to the standard deviation of the δ_s's across sites. Third, the bias depends on the amount of between-site variation in compliance relative to the magnitude of the average compliance across sites. Holding constant $\text{Corr}(\gamma_s, \delta_s)$, τ_δ, and F, the magnitude of the compliance-effect covariance bias is maximized when $|\text{CV}_\gamma| = 1$ (see online Appendix A4). As CV_γ approaches 0 (in which case the compliance is homogeneous across sites) or $\pm\infty$ (i.e., as the average compliance across sites goes to 0), the compliance-effect covariance bias term goes to 0. And fourth, the compliance-effect covariance bias is smaller when F is small. When the instruments are collectively strong, the bias due to between-site compliance-effect covariance is maximized.[4] Thus, compliance-effect covariance can lead to bias in the 2SLS estimator even with an arbitrarily strong set of instruments.

Each of the four factors influencing the compliance-effect covariance bias component is, in principle, estimable from the observed data (although estimation of τ_δ and $\text{Corr}(\gamma_s, \delta_s)$ will be complicated by FSB in the estimation of the δ_s's). The correlation between the first- and second-stage error terms is not estimable from the observed data however. When F is large, however, the contribution of FSB to the overall bias is negligible. This suggests that we may be able to devise a better estimator of δ—one that is less biased by compliance-effect covariance—than 2SLS, at least for the case where F is relatively large. In Section IV of this article, we develop two such estimators.

Equation 5a provides an approximation to the bias induced by the combination of finite within-site samples and compliance-effect covariance. However, Equation 5a does not describe the sampling variance of the 2SLS estimator in the presence of compliance and effect heterogeneity, compliance-effect covariance, and finite within-site samples. It is well known that 2SLS yields standard errors that are too small when there are many weak instruments, but these results have been developed under the assumption that δ_s is constant across sites (Angrist & Pischke, 2009; Chamberlain & Imbens, 2004). In the following section, we conduct a set of simulation analyses to describe the sampling variance of the OLS and 2SLS estimators in the presence of heterogeneous compliance and effect.

III. Simulation Analyses

This section presents results from a series of simulations conducted with three goals: (1) to test whether the 2SLS bias formula presented in Equation 5a is accurate (since it is based on an approximation) and to examine the extent of 2SLS bias that exists under a range of conditions; (2) to assess the sampling variation of the 2SLS estimator in the presence of compliance and effect heterogeneity and compliance-effect covariance; and (3) to compare the magnitude of the bias and the RMSE of the 2SLS estimator relative to the OLS estimator. To simplify matters, the within-site variance of the individual compliance and effect parameters are set to zero; therefore, these simulations focus on variation and covariance of γ_s and δ_s *across site*, not within site. The Appendix provides a more detailed description of the simulation setup.

Results of the simulations are shown in Table 1. In each panel of Table 1, one of the four key parameters that influence the bias and sampling variability of the 2SLS estimator—CV_γ, the expected first-stage *F*-statistic, the compliance-effect correlation, $Corr(\gamma_s, \delta_s)$, and the variance of the effect, τ_δ—is systematically manipulated while the other three are held constant (see the Appendix for details). Note that, except for Panel B, we set the *F*-statistic to 10. Columns 5 through 14 report the results obtained from 2,000 simulation samples drawn from a population of sites generated according to the parameter values shown in columns 1 through 4. In each case, the simulated data are generated based on a true effect of $\delta = 1$, so the bias reported in columns 5 and 6 of Table 1 can be interpreted as the ratio of the bias to the magnitude of the true effect.

Magnitude of the Estimated 2SLS Bias in the Presence of Compliance-Effect Covariance in Finite Samples

In Table 1, column 5 reports the predicted 2SLS bias as computed from Equation 5a. Column 6 reports the estimated 2SLS bias from the simulations (the difference between the average 2SLS estimate over the 2,000 simulations and the true effect). In each case, the estimated bias in column 6 is very close to that predicted by Equation 5a. As expected, the bias is larger when CV_γ is near 1; when *F* is small; when the correlation of γ_s and δ_s is large; and when the variance of δ_s is large. One key lesson from Table 1 is that 2SLS bias can be substantial, even when $F \geq 10$, particularly when the absolute value of the compliance-effect correlation is large or the variance of δ is large (see rows 11, 15, and 19).

Sampling Variability of the 2SLS Estimator in the Presence of Compliance-Effect Covariance in Finite Samples

Table 1 reports both the true sampling variation (column 7; the standard deviation of the 2SLS estimates of δ across the 2,000 simulation samples) and the average standard error reported by conventional 2SLS estimation algorithms

TABLE 1.
Estimated Bias and Root Mean Squared Error of Multiple-Site, Multiple-Instrument 2SLS Estimator

	Data-Generating Parameters				2SLS Estimator					OLS Estimator				
Case	CV_γ (1)	F (2)	Corr(γ_s,δ_s) (3)	SD (δ_s) (4)	Predicted Bias (5)	Estimated Bias (6)	True se($\hat{\delta}$) (7)	Average se($\hat{\delta}$) (8)	RMSE (9)	Predicted Bias (10)	Estimated Bias (11)	True se($\hat{\delta}$) (12)	Average se($\hat{\delta}$) (13)	RMSE (14)
Panel A: CV_γ varies														
1	0	10	0.25	1	0.050	0.051	0.173	0.064	0.180	0.479	0.483	0.142	0.013	0.503
2	0.2	10	0.25	1	0.137	0.137	0.173	0.061	0.221	0.482	0.487	0.139	0.013	0.506
3	1	10	0.25	1	0.275	0.283	0.223	0.061	0.361	0.489	0.494	0.139	0.013	0.513
4	5	10	0.25	1	0.137	0.151	0.256	0.062	0.297	0.482	0.488	0.139	0.013	0.507
5	∞	10	0.25	1	0.050	0.074	0.259	0.064	0.270	0.479	0.485	0.140	0.013	0.504
Panel B: Expected F-statistic varies														
6	1	2	0.25	1	0.375	0.387	0.243	0.135	0.457	0.499	0.504	0.139	0.013	0.523
7	1	5	0.25	1	0.300	0.309	0.229	0.086	0.385	0.495	0.500	0.139	0.013	0.519
8	1	10	0.25	1	0.275	0.283	0.223	0.061	0.361	0.489	0.494	0.139	0.013	0.513
9	1	26	0.25	1	0.260	0.267	0.220	0.039	0.346	0.472	0.478	0.139	0.013	0.497
10	1	101	0.25	1	0.252	0.259	0.218	0.021	0.339	0.416	0.423	0.146	0.013	0.447
Panel C: Corr(γ_s,δ_s) varies														
11	1	10	−0.75	1	−0.625	−0.603	0.240	0.078	0.649	0.445	0.446	0.145	0.013	0.469
12	1	10	−0.25	1	−0.175	−0.157	0.225	0.067	0.275	0.467	0.471	0.139	0.013	0.491
13	1	10	0	1	0.050	0.063	0.223	0.063	0.232	0.478	0.483	0.138	0.013	0.502
14	1	10	0.25	1	0.275	0.283	0.223	0.061	0.361	0.489	0.494	0.139	0.013	0.513
15	1	10	0.75	1	0.725	0.720	0.234	0.061	0.757	0.511	0.517	0.142	0.013	0.536
Panel D: SD (δ) varies														
16	1	10	0.25	0	0.050	0.051	0.045	0.044	0.068	0.478	0.479	0.009	0.010	0.479
17	1	10	0.25	0.2	0.095	0.096	0.061	0.044	0.114	0.480	0.482	0.028	0.012	0.483
18	1	10	0.25	1	0.275	0.283	0.223	0.061	0.361	0.489	0.494	0.139	0.013	0.513
19	1	10	0.25	5	1.175	1.221	1.101	0.234	1.644	0.533	0.558	0.696	0.013	0.892

Note: 2SLS = two-stage least squares; OLS = ordinary least squares; RMSE = root mean squared error.
Details of simulation in the Appendix. In each row, the following parameters are used: Each simulation data set has 50 sites, with 200 observations within site, 50% of which are assigned to the treatment condition. The variances of the first- and second-stage error terms are set to 1, and their correlation is set to .5. In column 5, the predicted bias is computed from Equation 5a; in column 10, the predicted bias is computed from Equation 4a. The RMSE in column 9 is computed as the square root of the sum of the squares of columns 6 and 7. The RMSE in column 14 is computed as the square root of the sum of the squares of columns 11 and 12.

(column 8). These conventional 2SLS-estimated standard errors are based on the assumption that δ_s is constant across sites. Equation B14 in Reardon and Raudenbush (in press), however, implies that the sampling variance of the 2SLS estimator depends on the variance of δ_s; assuming that $\tau_\delta = 0$ will lead one to underestimate the sampling variance of the 2SLS estimator. This is evident in comparing columns 7 and 8 in Table 1. The true sampling variance of the 2SLS estimates is generally much larger than that implied by the conventional 2SLS-estimated standard errors. Only in row 16, where τ_δ is set to zero, does the 2SLS standard error appropriately match the true sampling variance of the estimator. Note that this result is not merely due to the fact that, in overidentified 2SLS models, the estimated standard errors are often too small, especially when the instruments are collectively weak (Angrist & Pischke, 2009; Chamberlain & Imbens, 2004). Even in row 10, where the *F*-statistic is 101, the 2SLS-estimated standard error is one tenth of the true sampling standard deviation. We conclude that conventional 2SLS-estimated standard errors may substantially underestimate the sampling variance of the estimates when the mediator effect is heterogeneous.

Comparing 2SLS and OLS Estimators in the Presence of Compliance-Effect Covariance in Finite Samples

It is useful to compare the performance of the 2SLS estimator to the OLS estimator in the presence of compliance-effect covariance. Table 1 includes four columns that make this comparison possible: Columns 10 and 11 report the predicted OLS bias (based on Equation 4a) and the estimated OLS bias, respectively; column 12 reports the true OLS sampling variation (the standard deviation of the OLS estimates across the 2,000 simulation samples in each case); column 13 reports the average reported OLS-estimated standard error across the 2,000 samples; and column 14 shows the RMSE for OLS (the square root of the sum of the squares of columns 11 and 12). These results lead to three observations: First, for the range of the parameters tested, OLS bias tends to be larger than 2SLS bias. Second, the average OLS-estimated standard error substantially underestimates the true variability of the OLS estimator (unless $\tau_\delta = 0$), which tends to be smaller than the true variability of the 2SLS estimator. Finally, the RMSE for the OLS estimator tends to be larger than the RMSE of the 2SLS estimator because the OLS bias is generally larger than the 2SLS bias even though OLS estimates are more precise than 2SLS. Note that this does not apply to cases where the 2SLS bias is larger than the OLS bias due to a large compliance-effect covariance (e.g., rows 11, 15, and 19).

We draw three primary conclusions from the described simulation analysis. First, Equations 4a and 5a provide good approximations of the 2SLS and OLS biases in finite samples and in the presence of site-level compliance-effect covariance. Second, even when the instruments are collectively strong, conventional 2SLS-estimated standard errors substantially underestimate sampling variance

when the mediator effect is heterogeneous across sites. Third, unless compliance-effect covariance bias is large, the 2SLS estimator generally has less bias but larger sampling variance than the OLS estimator; consequently, the RMSE for the OLS estimator tends to be larger than that of the 2SLS estimator. Although the presence of compliance-effect covariance leads to some bias, it may generally not be so large as to render 2SLS less desirable than OLS.

IV. Two Bias-Corrected Multisite Single Mediator IV Estimators

In Section II, we demonstrated that 2SLS yields biased estimates of the average effect of M when there is between-site compliance-effect covariance, even if F is arbitrarily large. As we suggested there, however, because the magnitude of compliance-effect covariance bias may be estimable from the observed data under certain conditions, it may be possible to develop a method of correcting the 2SLS estimates to eliminate this bias.

To build some intuition regarding our approach, consider the hypothetical data described in Figure 1. Each of the panels on the left side of the figure shows a hypothetical relationship between δ_s and γ_s. In each case, δ (the average value of δ_s across sites) equals 1. Likewise, in each case, the average compliance across sites equals 1, and both γ_s and δ_s have a variance of 1. This implies that $CV_\gamma = 1$, so these figures correspond to cases in which compliance-effect covariance bias is maximized (for a given value of F, $\text{Corr}(\gamma_s, \delta_s)$, and τ_δ). The three figures on the left side differ only in the correlation between δ_s and γ_s, ranging from $\text{Corr}(\delta_s, \gamma_s) = -0.50$ to $\text{Corr}(\delta_s, \gamma_s) = +0.50$.

Under the assumptions that treatment affects the outcome only through the mediator (exclusion restriction) and there is no within-site compliance-effect covariance, the average intent-to-treat effect on the outcome within a site s will be $\beta_s = \gamma_s\delta_s$. The figures on the right side plot these computed intention-to-treat (ITT) effects against the γ_s's. In practice, we can estimate the β_s's and the γ_s's, so we can readily produce figures of the type shown here. Note that a nonzero correlation between γ_s and δ_s will produce a figure on the right that shows a nonlinear association between β_s and γ_s. This is evident in the quadratic-fitted curves in the right-hand figures. Thus, nonlinearity in the observed relationship between β_s and γ_s is informative regarding the extent of compliance-effect covariance across sites, and so may be useful in developing a bias-corrected estimator.

2SLS is equivalent to a linear regression of β_s on γ_s (albeit with no intercept, as the exclusion restriction requires that $\beta_s = 0$ when $\gamma_s = 0$), weighting each site by its sample size and the variance of the instrument within each site (Raudenbush et al., 2012; Reardon & Raudenbush, in press).[5] The slope of this line is the 2SLS IV estimate of δ. Recall that the average value of δ_s is 1, so an unbiased estimate would yield a slope of 1, as shown by the solid line in the figures. The results of the 2SLS regression are shown by the dashed line. Note that when $\text{Corr}(\delta_s, \gamma_s) > 0$, the slope of the fitted line is substantially greater than 1; when

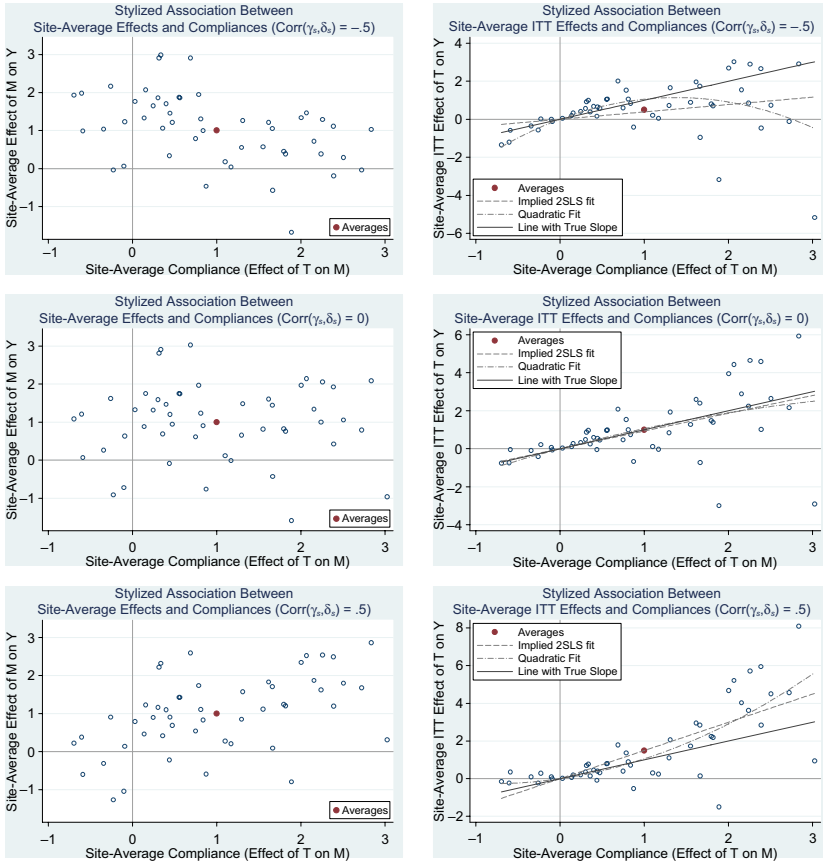

FIGURE 1. *Stylized association between site-average effect and compliance and site-average intention-to-treat (ITT) effect and compliance.*

$\text{Corr}(\delta_s, \gamma_s) < 0$, the slope of the line is substantially less than 1. The reason for this is that the sites where γ_s is largest in magnitude (farthest from 0) have more leverage in the regression; the correlation between γ_s and δ_s means that these sites also have larger (or smaller) than average δ_s's, leading to biased estimates.

Two Bias-Corrected Estimators

We now develop two bias-corrected IV estimators. First, assume that the association between γ_s and δ_s is linear:

$$\delta_s = \alpha_0 + \alpha_1 \gamma_s + \nu_s, \quad \nu_s \sim N\left[0, \sigma_\nu^2\right]. \tag{7}$$

Note that this assumption is weaker than the assumption that $\text{Cov}(\gamma_s, \delta_s) = 0$. We can, in principle, relax the linearity assumption further, and allow the relationship between γ_s and δ_s to be described by some higher order polynomial. Equation 8 would then include a set of terms involving the expected values of the higher order powers of γ_s. This would result in a higher order regression model in Equation 12.

Taking the expectation of both sides of Equation 7 yields

$$\begin{aligned} E[\delta_s] &= \alpha_0 + \alpha_1 E[\gamma_s] \\ \delta &= \alpha_0 + \alpha_1 \gamma. \end{aligned} \tag{8}$$

Equation 8 suggests a bias-corrected estimator for δ. Specifically, if we can estimate α_0, α_1, and γ, we can estimate δ as

$$\hat{\delta}^{bc} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\gamma}. \tag{9}$$

We can construct a second bias-corrected estimator by directly estimating the 2SLS compliance-effect covariance bias and subtracting it from the 2SLS estimate. Note that Equation 7 implies that $\alpha_1 = \tau_{\gamma\delta}/\tau_\gamma$; the compliance-effect covariance bias (from Equation 5a) is therefore

$$\frac{2\gamma\alpha_1\tau_\gamma}{\gamma^2 + \tau_\gamma}\left(\frac{F-1}{F}\right). \tag{10}$$

Thus, if we could estimate F, α_1, γ, and τ_γ, we can construct a plug-in bias-corrected estimator:

$$\hat{\delta}^{pi} = \hat{\delta}^{2\text{SLS}} - \frac{2\hat{\gamma}\hat{\alpha}_1\hat{\tau}_\gamma}{\hat{\gamma}^2 + \hat{\tau}_\gamma}\left(\frac{\widehat{F}-1}{\widehat{F}}\right). \tag{11}$$

To construct $\hat{\delta}^{bc}$ and $\hat{\delta}^{pi}$, we must estimate F, α_0, α_1, γ, and τ_γ First, we can estimate γ, τ_γ, and F from the following random-coefficients model:[6]

$$\begin{aligned} M_i &= \Lambda_s + \gamma_s T_i + e_i \\ \begin{pmatrix} \Lambda_s \\ \gamma_s \end{pmatrix} &\sim N\left[\begin{pmatrix} \Lambda \\ \gamma \end{pmatrix}, \begin{pmatrix} \tau_\Lambda & \tau_{\gamma\Lambda} \\ \tau_{\gamma\Lambda} & \tau_\gamma \end{pmatrix}\right]. \end{aligned} \tag{12}$$

Now, we note that

$$\begin{aligned} \beta_s &= E[B|S=s] \\ &= E[\Gamma\Delta|S=s] \\ &= E[\Gamma|S=s] \times E[\Delta|S=s] + \text{Cov}(\Gamma\Delta)|S=s \\ &= \gamma_s \times \delta_s + \text{Cov}(\Gamma\Delta)|S=s. \end{aligned} \tag{13}$$

Given the assumption of no within-site compliance-effect covariance, substituting Equation 7 into 13 yields

$$\begin{aligned}\beta_s &= \gamma_s \times \delta_s \\ &= \gamma_s(\alpha_0 + \alpha_1\gamma_s + \nu_s) \\ &= \alpha_0\gamma_s + \alpha_1\gamma_s^2 + \gamma_s\nu_s, \quad \nu_s \sim N\left[0, \sigma_\nu^2\right].\end{aligned} \tag{14}$$

In other words, under the assumption that δ_s is linearly related to γ_s, β_s can be written as a quadratic function of γ_s, passing through the origin, with a heteroscedastic error term. The parameters α_0 and α_1 can be estimated by fitting this model to the $\hat{\beta}_s$'s and $\hat{\gamma}_s$'s.

Although the assumption that T is ignorably assigned within sites ensures that we can obtain unbiased estimates of the β_s's and γ_s's, two factors will complicate the estimation of α_0 and α_1 from the observed data. First, we do not observe β_s and γ_s; rather, we estimate them and so observe $\hat{\beta}_s = \beta_s + b_s$ and $\hat{\gamma}_s = \gamma_s + g_s$. Regressing $\hat{\beta}_s$ on $\hat{\gamma}_s$ and $\hat{\gamma}_s^2$ will yield biased estimates of α_0 and α_1 because of the error in the $\hat{\gamma}$'s. Second, in finite samples, the correlation between e and u (the errors in the first- and second-stage equations) will induce a correlation between b_s and g_s, as will $\delta \neq 0$ (see Equation A3.5 in online Appendix A3); this will induce bias in the estimates of α_0 and α_1.

We can correct the first problem by regressing the $\hat{\beta}_s$'s on shrunken estimates of γ_s and γ_s^2. In online Appendix A5 we show that

$$E\left[\hat{\beta}_s|\hat{\gamma}_s\right] = \alpha_0\gamma_s^* + \alpha_1\gamma_s^{2*} + \mathrm{Cov}(b_s, g_s)\frac{\lambda(\hat{\gamma}_s - \gamma)}{\tau_\gamma}, \tag{15}$$

where $\lambda = \tau_\gamma/(\tau_\gamma + \tau_g)$ is the reliability of the $\hat{\gamma}_s$'s; $\gamma_s^* = E\left[\gamma_s|\hat{\gamma}_s\right] = \lambda\hat{\gamma}_s + (1-\lambda)\gamma$; and $\gamma_s^{2*} = E\left[\gamma_s^2|\hat{\gamma}_s\right] = \gamma_s^{*2} + \tau_\gamma(1-\lambda)$. When F is large and CV_γ is not small, the expected value of the final term in Equation 15 will be small.[7] This suggests we can regress the $\hat{\beta}_s$'s on γ_s^* and γ_s^{2*} (with no intercept) to estimate α_0 and α_1. Given the estimates $\hat{\gamma}$, $\hat{\tau}_\gamma$, $\hat{F}$, $\hat{\alpha}_0$, and $\hat{\alpha}_1$, we can then compute $\hat{\delta}^{bc}$ and $\hat{\delta}^{pi}$ from Equations 9 and 11. If we have large samples within sites, we can estimate the β_s's and γ_s's very reliably, which will lead to precise estimates of γ, τ_γ, α_0, and α_1, and thus, to precise estimates of δ. Note that $\hat{\delta}^{pi}$ and $\hat{\delta}^{bc}$ rely on the same basic information (both use $\hat{\gamma}$, $\hat{\tau}_\gamma$, and $\hat{\alpha}_1$; $\hat{\delta}^{bc}$ also uses $\hat{\alpha}_0$, however), but in different ways, suggesting that they may perform somewhat differently under different conditions.

Standard Errors for $\hat{\delta}^{\mathrm{bc}}$ and $\hat{\delta}^{\mathrm{pi}}$

We compute standard errors for $\hat{\delta}^{bc}$ and $\hat{\delta}^{pi}$ via bootstrapping. Specifically, we (a) draw a sample of K sites, with replacement, from the original sample of sites; (b) draw a sample of $p \times n$ treatment and $(1-p)n$ control cases, with

replacement, separately in each resampled site; (c) estimate $\hat{\delta}^{bc}$ and $\hat{\delta}^{pi}$ from this new sample as described in Equations 9 and 11; (d) repeat Steps (a) through (c) many times (we use 500 draws in the simulations described later); and (e) use the variances of the estimates from these repeated draws as estimates of the sampling variances of the $\hat{\delta}$'s.

V. Simulation Analyses

We assess the performance of the two bias-corrected IV estimators described in Section IV using a set of simulations, comparing the results based on the new estimators with those from 2SLS. The Appendix describes the simulation setup in detail. We vary three parameters—the coefficient of variation for compliance (CV_γ), the expected *F*-statistic, and the compliance-effect correlation—across simulations.

Table 2 presents the estimated bias, sampling variation, estimated standard error, and RMSE of the two bias-corrected estimators for a range of simulated populations. Columns 1 through 3 report the parameters used in each simulation; columns 4 through 11 report the estimation results for the two bias-corrected IV estimators. For comparison, columns 12 through 15 report the corresponding 2SLS bias, standard error, and RMSE.

Bias of the Bias-Corrected IV Estimators in the Presence of Compliance-Effect Covariance in Finite Samples

Columns 4 and 8 in Table 2 present the estimated bias for the two bias-corrected estimators across 2,000 simulation iterations.[8] Panel A indicates that the magnitude of the estimated bias of both bias-corrected estimators reaches its *minimum* value when CV_γ equals 1, other things being equal. As CV_γ deviates from 1, the absolute value of estimated bias increases.[9] Thus, the bias-corrected estimators are most effective at eliminating bias when CV_γ is near 1. This is in stark contrast with the pattern observed in panel A of Table 1 and in columns 12 through 15 of Table 2, which show that bias in 2SLS exhibits an inverse "U" shape that reaches its maximum value when CV_γ is 1 and diminishes steadily as CV_γ starts deviates from 1 in either direction. Panel A also indicates that the bias-corrected estimator exhibits less bias than the plug-in estimator when $CV_\gamma < 1$, and more bias than the plug-in estimator when $CV_\gamma \geq 1$.

Panel B suggests that both bias-corrected estimators do a good job eliminating bias when the first-stage *F*-statistic is large. A comparison between columns 4, 8, and 12 indicates that, when *F* is extremely small, the absolute value of bias of the bias-corrected estimators is similar to that of 2SLS. As *F* increases, the magnitude of the bias shown in columns 4 and 8 decreases both in absolute terms and as a proportion of 2SLS bias.

TABLE 2.

Estimated Bias and RMSE of Bias-Corrected IV Estimator and Multiple-Site, Multiple-Instrument 2SLS IV Estimator

	Data-Generating Parameters			Bias-Corrected IV Estimator				Plug-In Bias-Corrected IV Estimator				2SLS Estimator			
Case	CV_γ (1)	F (2)	Corr(γ_s,δ_s) (3)	Estimated Bias (4)	True se($\hat{\delta}$) (5)	Average se($\hat{\delta}$) (6)	RMSE (7)	Estimated Bias (8)	True se($\hat{\delta}$) (9)	Average se($\hat{\delta}$) (10)	RMSE (11)	Estimated Bias (12)	True se($\hat{\delta}$) (13)	Average se($\hat{\delta}$) (14)	RMSE (15)
Panel A: CV_γ varies															
1	0	10	0.25	−0.178	0.144	0.160	0.228	−0.258	0.133	0.160	0.290	0.051	0.173	0.064	0.180
2	0.2	10	0.25	−0.182	0.152	0.170	0.237	−0.265	0.142	0.150	0.301	0.137	0.173	0.061	0.221
3	1	10	0.25	0.083	0.278	0.270	0.290	−0.007	0.245	0.190	0.245	0.283	0.223	0.061	0.361
4	5	10	0.25	0.189	0.248	0.220	0.312	0.072	0.238	0.210	0.249	0.151	0.256	0.062	0.297
5	∞	10	0.25	0.193	0.251	0.230	0.317	0.073	0.241	0.220	0.252	0.074	0.259	0.064	0.270
1	0	26	0.25	−0.062	0.140	0.148	0.153	−0.098	0.135	0.140	0.167	0.020	0.157	0.040	0.158
2	0.2	26	0.25	−0.062	0.142	0.152	0.155	−0.100	0.138	0.150	0.171	0.116	0.159	0.039	0.196
3	1	26	0.25	0.039	0.230	0.223	0.233	0.002	0.217	0.200	0.217	0.267	0.220	0.039	0.346
4	5	26	0.25	0.080	0.227	0.213	0.241	0.040	0.236	0.220	0.239	0.129	0.257	0.040	0.287
5	∞	26	0.25	0.082	0.230	0.215	0.244	0.041	0.240	0.220	0.244	0.043	0.260	0.041	0.264
Panel B: Expected F-statistic varies															
6	1	2	0.25	−0.349	2.664	0.774	2.687	−0.346	1.052	0.270	1.107	0.387	0.243	0.135	0.457
7	1	5	0.25	0.114	0.435	0.359	0.450	−0.044	0.337	0.270	0.340	0.309	0.229	0.086	0.385
8	1	10	0.25	0.083	0.278	0.270	0.290	−0.007	0.245	0.190	0.245	0.283	0.223	0.061	0.361
9	1	26	0.25	0.039	0.230	0.223	0.233	0.002	0.217	0.200	0.217	0.267	0.220	0.039	0.346
10	1	101	0.25	0.014	0.215	0.210	0.215	0.003	0.210	0.200	0.210	0.259	0.218	0.021	0.339
Panel C: Corr(γ_s,δ_s) varies															
11	1	26	0.00	0.058	0.236	0.229	0.243	0.021	0.221	0.210	0.222	0.030	0.223	0.040	0.232
12	1	26	0.25	0.039	0.230	0.223	0.233	0.002	0.217	0.200	0.217	0.270	0.223	0.040	0.361
13	1	26	0.75	−0.001	0.189	0.189	0.189	−0.039	0.209	0.190	0.213	0.730	0.234	0.040	0.757

Note: 2SLS = two-stage least squares; RMSE = root mean squared error; IV = instrumental variables.

Details of simulation in the Appendix. In each row, $\delta = 1$ and $SD(\delta) = 1$. All additional parameters are set as described in Table 1. Columns 5 and 9 report the standard deviation of the distribution of estimates of δ over 2,000 samples. Column 6 reports the average bootstrapped standard error (see text for description of bootstrapping procedure) over 100 samples (bootstrapped standard errors were computed for only 100 iterations due to computational time). The RMSE in columns 7, 11, and 15 are computed as described in Table 1.

Panel C shows that, for cases examined here, bias in the bias-corrected estimator decreases as $\text{Corr}(\gamma_s, \delta_s)$ increases, other things being equal. Bias in the plug-in estimator appears slightly larger in the case where $\text{Corr}(\gamma_s, \delta_s)$ is larger than where it is moderate in size, though it is still very small compared to the bias in the 2SLS estimator. Note that compliance-effect bias (CEB) in 2SLS or OLS increases with $\text{Corr}(\gamma_s, \delta_s)$. So results in this panel indicate that, when CV_γ is 1 and the *F*-statistic is fairly large (e.g., $F = 26$), both bias-corrected estimators performs very well when they are needed the most—when the CEB is large.

Sampling Variability of the Bias-Corrected IV Estimators in the Presence of Compliance-Effect Covariance in Finite Samples

Columns 5 and 9 report the true sampling variation (the standard deviation of the estimates across the 2,000 simulation samples) of the two bias-corrected estimators, while columns 6 and 10 report the average bootstrapped standard error of the estimates, for each scenario. In general, except when *F* is very small, the sampling variance of both bias-corrected estimators is roughly similar to that of the 2SLS estimates. This suggests that the bias correction does not come at any significant loss of precision compared to 2SLS (of course, the sampling variances of the bias-corrected estimators and of 2SLS are much larger than the conventional 2SLS-estimated standard errors, as shown in column 14). Moreover, the bootstrapped standard errors for the bias-corrected estimators are very close to the true standard errors, except when *F* is very small.

Comparing the Bias-Corrected IV Estimators to the 2SLS and OLS Estimators in the Presence of Compliance-Effect Covariance in Finite Samples

Figure 2 compares the estimated bias and RMSE from the 2SLS and bias-corrected IV estimators under a variety of conditions. The horizontal axis in each graph indicates the first-stage *F*-statistic and the vertical axis either the bias (left panel of figures) or RMSE (right panel). We present separate graphs for CV_γ values of 1.0, 0.2, and 0.[10] In each case, $\text{Corr}(\gamma_s, \delta_s)$ is fixed at 0.25.

In each of the graphs on the left panel, the area below the 2SLS bias line is decomposed into two parts: The light gray area on top represents the amount of CEB in the 2SLS estimator and the dark gray area at the bottom represents the FSB component of the 2SLS estimator. This decomposition is based on Equation 5a and the sum of these two components closely tracks the estimated 2SLS bias (the sum does not exactly track the bias as the decomposition is an approximation). These three graphs illustrate that the relative bias of 2SLS and the bias-corrected estimators depends both on CV_γ and the first-stage *F*-statistic. As expected, the bias-corrected estimators reduce 2SLS bias the most when 2SLS CEB is large relative to the 2SLS FSB.

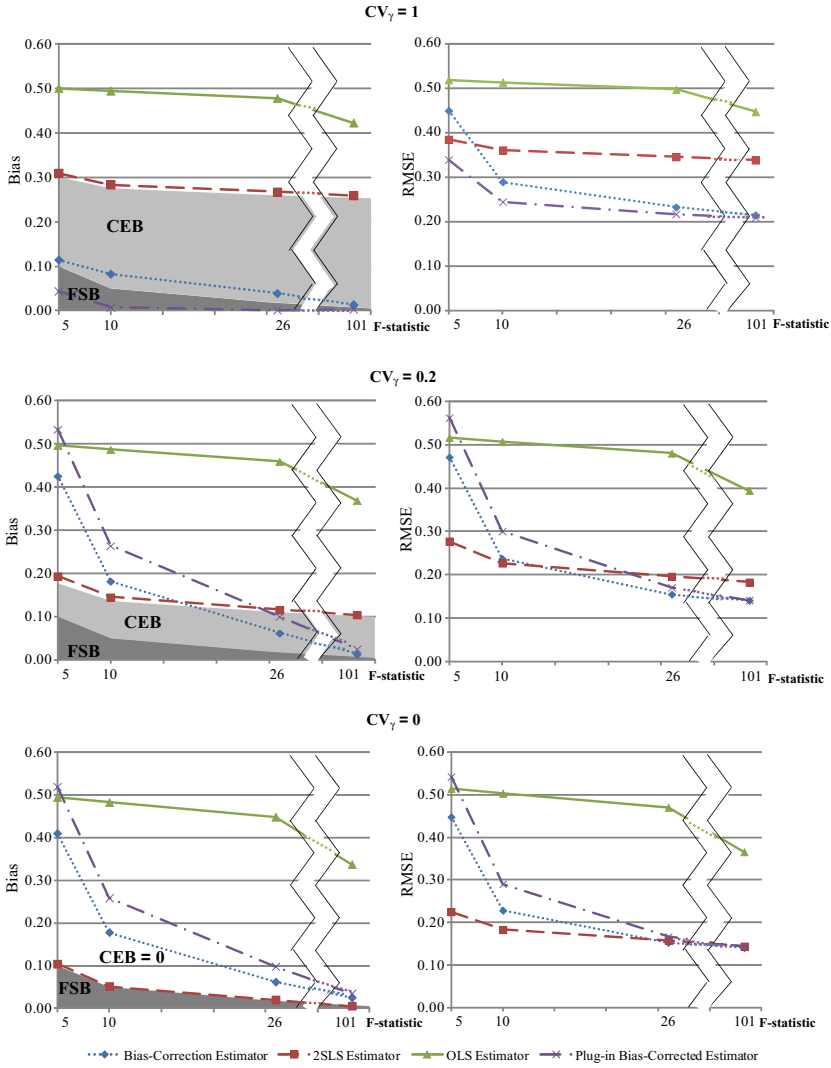

FIGURE 2. *Bias and RMSE of four estimators by* F*-statistic and CV$_\gamma$, when Corr (γ_S, δ_S) = 0.25. RMSE = root mean squared error.*

Specifically, when CV$_\gamma$ is 1, the bias-corrected estimators always have smaller bias than the 2SLS estimator, regardless of the first-stage *F*-statistic (top graph). This is not surprising since, for any given *F*-statistic, the bias-corrected estimators have minimum bias when CV$_\gamma$ is 1, while 2SLS bias is maximized at this point. The dotted line closely tracks the FSB area (in dark gray), indicating

that, in this case, the bias-corrected estimators are very successful in eliminating almost all of the CEB in the 2SLS estimator, regardless of the *F*-statistic.

When CV_γ is different from 1 but does not lie in the extremes (i.e., $CV_\gamma = 0.2$), the bias-corrected estimators can still produce a smaller bias than the 2SLS method if the *F*-statistic is greater than 10 (middle graph). As CV_γ continues to deviate from 1 and reaches the extreme of zero (i.e., when γ_s does not vary across sites), the bias in the bias-corrected estimators approaches the bias in the 2SLS estimator as the *F*-statistic increases, but the 2SLS estimator produces the smallest bias among the four methods for all *F*-statistics presented here (bottom graph). This is not surprising since in this case, there is no CEB in the 2SLS estimator (the first term in Equation 5a is zero), therefore there is nothing for the alternative method to correct for.

The three graphs on the right side of Figure 2 compare the RMSE of these four estimators. The layout for these graphs is the same as that for the graphs on the left side except that the vertical axis now represents the RMSE instead of the bias. These three graphs show that the RMSE for the bias-corrected estimators is generally larger than that for the 2SLS estimator when *F* is small (less than 10), but decreases faster as *F* increases than does the RMSE of the 2SLS estimator. As a result, the bias-corrected IV estimators have the smallest RMSE when *F* is above some threshold, though this threshold depends on CV_γ—it is the smallest when CV_γ is 1 (top graph) and becomes larger as CV_γ deviates from 1 (middle and bottom graph).

Figure 3 provides similar comparisons of the magnitude of bias and RMSE among the three estimators as a function of the compliance-effect correlation. In these figures, the *F*-statistic is set to a value of 26, and the horizontal axis indicates values of $\text{Corr}(\gamma_s, \delta_s)$. All other attributes of the graph are the same as in Figure 2.[11]

Similar to Figure 2, the three graphs on the left side of Figure 3 show that when $CV_\gamma = 1$, the bias-corrected estimators work well in eliminating the CEB in the 2SLS bias, especially when the CEB is large (top graph). When CV_γ deviates somewhat from 1, the bias-corrected estimators eliminate some, but not all of the CEB (middle graph). When there is no CEB in the 2SLS estimator (either because $\text{Corr}(\gamma_s, \delta_s) = 0$ or $CV_\gamma = 0$), the bias in the 2SLS estimator is smaller than that of the bias-corrected estimator. Nonetheless, as the three graphs on the right side of Figure 3 show that, across all cases examined in this figure, the RMSE of the bias-corrected estimator is always smaller or equal to that of 2SLS, even when there is no compliance-effect covariance bias. The RMSE of the plug-in estimator is similar in magnitude, although it is sometimes larger than that of 2SLS. This suggests that the bias-corrected IV estimator may be generally preferable to 2SLS as long as *F* is modestly large (recall that it is 26 in the figures here) and $CV_\gamma \leq 1$.

It is clear that the combination of CV_γ, the *F*-statistic, and $\text{Corr}(\gamma_s, \delta_s)$ affect the performance of the bias-corrected IV estimators relative to that of the 2SLS

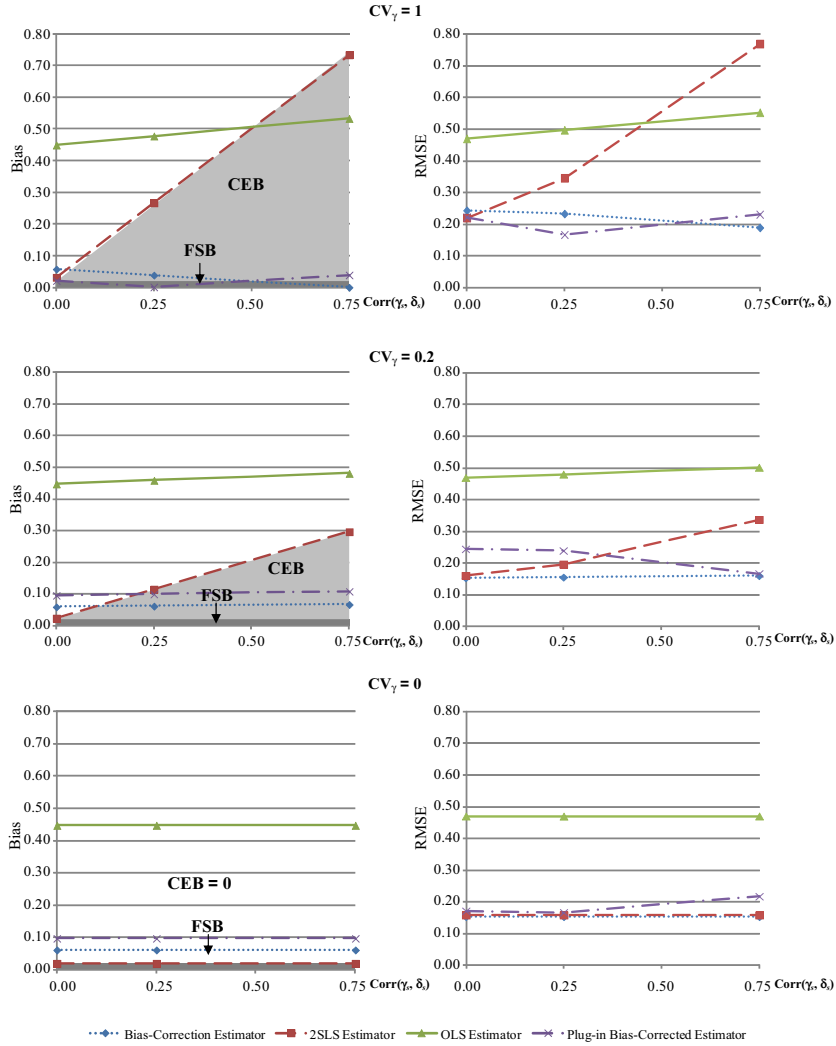

FIGURE 3. *Bias and RMSE of four estimators by Corr(γ_S, δ_S) and CV_γ, when* F*-statistic = 26. RMSE = root mean squared error.*

estimator. In general, when the *F*-statistic is greater than 10, the bias-corrected IV estimators outperform the 2SLS estimator both in terms of bias and RMSE under a wide range of conditions. This is especially true when CV_γ does not deviate from 1 too much and when $\mathrm{Corr}(\gamma_s, \delta_s)$ is not very close to zero. When the *F*-statistic is less than 10, the bias-corrected estimators generally perform worse than 2SLS. However, because IV methods should generally not be used when the

F-statistic is less than 10 (see, e.g., Stock & Yogo, 2005), this is not a particularly useful comparison.

VI. Empirical Examples

We now apply 2SLS and the bias-corrected IV estimators to a reanalysis of data from two studies: (1) the Tennessee class size experiment, Project STAR (e.g., Finn & Achilles, 1990) and (2) the federal RF Impact study described earlier. For both examples, we estimate the relationship between a hypothesized mediator and an outcome using OLS, 2SLS, and the bias-corrected estimator. However, the examples represent two very different study designs. Project STAR randomly assigned a large number of individual students to treatment status in a large number of sites (schools), whereas the RF study examined student outcomes for a small number of schools that were assigned to treatment or control status in a small number of sites. The two examples also differ in terms of the factors that influence the effectiveness of our bias-corrected estimator: (1) the strength of their instruments, (2) their cross-site variation in compliance, and (3) their cross-site correlation between compliance and mediator effects. In addition, the RF study was a cluster-randomized trial: Schools, rather than students, were randomly assigned to treatment conditions. We take this clustering into account in our analyses, but do not spend time discussing the clustering issue, as it is orthogonal to the key issues of identification and bias that we focus on in this article.

Project STAR

Project STAR randomized approximately 5,900 entering kindergarten students at 79 elementary schools to either a small class (13–17 students) or a regular-sized class (22–26 students; Krueger, 1999, and Nye, Hedges, & Konstantopoulos, 2000). Students assigned to a regular-sized class were further randomly assigned to classes with or without a classroom aide. Because previous analyses found no difference in student outcomes for students in regular-sized classrooms with or without an aide (Krueger, 1999), we combine these two groups into a single regular-size classroom group.

The mediator of interest for us is actual class size, which differs from assigned class size because some students assigned to small classes ended up in classes with 18 or more students, and some assigned to regular classes had fewer than 22 in their class. Note that this mediator is an interval-scaled, multivalued variable rather than a binary "compliance" indicator. Therefore, this example is not simply a case where we are interested in adjusting the experimental estimates for noncompliance, but rather are interested in estimating the effect of a one-unit change in class size. As we show below, actual class size (and the effect of being assigned to a small class) varies significantly among students, even among those assigned to the same treatment condition. We use 79 instruments—a zero/one indicator for assignment to a small class interacted with a zero/one indicator for

TABLE 3.
Estimated Mediator Effects Using Empirical Data

	Project STAR		Reading First	
	Math	Reading	18 Blocks	36 Blocks
OLS estimator				
δ	−1.039**	−0.718**	0.037	0.122
Bootstrapped *SE*(δ)	(0.340)	(0.230)	(n.a.)	(n.a.)
2SLS estimator				
Δ	−1.114**	−0.714**	0.397	0.387
Bootstrapped *SE*(δ)	(0.350)	(0.230)	(n.a.)	(n.a.)
Observable/estimable parameters				
F-statistic	1,082.1	1,071.5	17.7	8.2
τ_γ	3.45	3.47	63.25	68.30
γ	−7.25	−7.26	10.47	10.45
CV_γ	0.26	0.26	0.76	0.79
Estimated Corr(γ_s, δ_s)	−0.240	−0.357	0.216	−0.009
Estimated 2SLS compliance-effect covariance bias	0.279	0.256	0.143	−0.005
Estimates from quadratic regression				
α_0	−3.583*	−3.025**	0.157	0.491
SE(α_0)	(1.546)	(0.959)	(1.025)	(0.783)
α_1	-0.312^{+}	−0.285*	0.020	−0.001
SE(α_1)	(0.187)	(0.116)	(0.060)	(0.045)
Bias-corrected estimator				
δ	−1.319**	−0.957***	0.365	0.484
Bootstrapped *SE*(δ)	(0.420)	(0.260)	(n.a.)	(n.a.)
Plug-in bias-corrected estimator				
δ	−1.392**	−0.969***	0.254	0.127
Bootstrapped *SE*(δ)	(0.418)	(0.269)	(n.a.)	(n.a.)
N (sites/blocks)	79	79	18	36
N (observations)	5,871	5,789	248	248

Note: STAR = Student–Teacher Achievement Ratio; OLS = ordinary least squares; 2SLS = two-stage least squares.
Estimated compliance-effect covariance bias computed from Equation 5a. Bootstrapped standard errors computed as described in text.
$+p < .10$. $*p < .05$. $**p < .01$. $***p < .001$.

each school. We use OLS, 2SLS with 79 instruments, and the two bias-corrected IV estimators with 79 instruments to estimate the effect of actual class size on student math and reading achievement at the end of the kindergarten year for students who were randomized when they entered kindergarten.

The left-hand panel of Table 3 summarizes the results of our reanalysis of the STAR data. We begin by considering the OLS and 2SLS estimates of the effects

of class size. The OLS estimates in Table 3 indicate that, on average, reducing the size of a kindergarten class by one student *increases* math achievement by 1.04 scale score points and *increases* reading achievement by 0.72 scale score points. The corresponding 2SLS estimates are 1.11 points in math and 0.71 points in reading, estimates that are very close to the OLS results. This similarity is likely because in Project STAR a very large proportion of the variance in class size was determined by random assignment, leaving little endogenous variation in class size to produce bias. Hence, unlike many mediators that vary naturally across individuals in a study sample, and thus may be correlated with their unobserved characteristics, Project STAR does not appear to have a substantial endogeneity problem.

Prior to estimating the effects of class size using the bias-corrected 2SLS estimators, it is useful to assess the potential compliance-effect covariance bias that might be present in the 2SLS estimates. To do so, we examine the F-statistic and estimate CV_γ, τ_δ, and $\mathrm{Corr}(\gamma, \delta)$ to determine whether, based on our simulations reported in Table 2 and Figures 2 and 3, we expect the bias-corrected estimators to outperform 2SLS. For both math and reading, $\mathrm{CV}_\gamma \approx 0.25$ and $F > 1,000$; the large F-statistic reflects the facts that variation in class size is largely due to randomization and that the average sample per school is substantial. Using the methods described in Raudenbush et al. (2012) and in online Appendix C, we estimate $\tau_\delta \approx 3.5$ for both math and reading, and $\mathrm{Corr}(\gamma, \delta) = -0.24$ and -0.36 in math and reading, respectively. These values suggest that the bias-corrected estimators should perform extremely well. Based on Figure 2, when $\mathrm{CV}_\gamma = 0.2$ and $\mathrm{Corr}(\gamma, \delta) = 0.25$, both the bias-corrected estimators are substantially less biased and have smaller RMSE when F is 100. Given that F is even larger in the STAR example (and given that 2SLS bias does not decline significantly after F is above 10), we prefer the bias-corrected IV estimates for these STAR analyses. Based on these values, Equation 5a implies that the compliance-effect covariance bias in the 2SLS estimator is roughly 0.27 in both math and reading; this is a moderate amount of bias relative to the 2SLS effect estimates of -1.11 and -0.71.

The two bias-corrected IV estimates (reported at the bottom of Table 3) are larger (18%–35% larger, in fact) than their 2SLS counterparts. They imply that reducing the size of a kindergarten class by one student increases average student achievement by 1.32 or 1.39 scale score points for math and 0.96 or 0.97 scale score points for reading (depending which of the two bias-corrected estimators we use). Expressed as effect sizes, these results imply a roughly 0.03 standard deviation increase in test scores per student of class size reduction. Note, however, that the standard errors of the bias-corrected estimates are 15%–20% larger than the 2SLS- and OLS-estimated standard errors and that the confidence intervals for the 2SLS, OLS, and bias-corrected estimates overlap considerably. For Project STAR, where variation in the mediator was mainly induced by randomization (and thus mainly exogenous) and where there are numerous randomized individuals

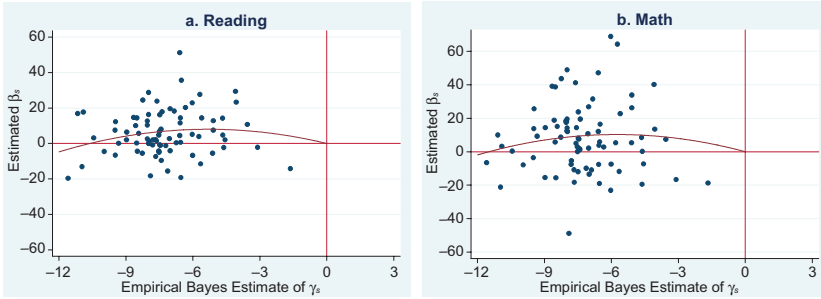

FIGURE 4. *Relationship between reduced form OLS estimates of β_s and empirical Bayes estimates of γ_s for each school in the Tennessee STAR sample for kindergarten reading and math test scores. OLS = ordinary least squares; STAR = Student–Teacher Achievement Ratio.*

per block and numerous blocks, the four estimation approaches yield roughly comparable point estimates and statistical inferences. Nonetheless, although our conclusions about the effectiveness of reducing class sizes may not change much depending on which the estimator we use in this case, the values of CV_γ, F, and $Corr(\gamma_s, \delta_s)$ and the simulations in Section V suggest that the two bias-corrected estimates are to be preferred to the 2SLS or OLS estimates in this example. As Figure 2 shows, when $CV_\gamma \approx 0.2$ and $Corr(\gamma_s, \delta_s) \approx 0.25$ and $F \geq 100$, the two bias-corrected estimators have very similar bias and RMSE; we have no clear way to choose between them in this case (nor do we need to, as they yield very similar estimates).

Another potential way to assess the impact of compliance-effect covariance bias is to examine the estimates of α_1. Because these estimates for Project STAR are statistically significant (at least in the case of reading) they provide reliable evidence of a true departure from linearity in the relationship between the effect of randomization on student achievement (β) and the effect of randomization on class size (γ). This departure from linearity implies the presence of compliance-effect covariance bias.

To help visualize this relationship, Figure 4 presents a graph of the reduced form OLS estimates of β_s against the Empirical Bayes estimates of γ_s for each school in the sample. Superimposed on this scatter plot is the estimated quadratic relationship implied by the estimates of α_0 and α_1 in Table 3. Because it is difficult to see a pattern in the plotted points, consider what is implied by the fitted curve. Sites in which there was a greater reduction in class size as a result of treatment assignment have, on average, a larger increase in test scores as a result of treatment assignment, but this association does not appear to be linear. This nonlinearity implies a covariance between the site-average compliance levels and site-average effects—a unit change in class size appears to effect test scores the

most, on average, in the schools where random assignment induced a smaller change in class size. This might result from a nonlinearity in the underlying relationship between class size and achievement.

RF Impact Study

The RF Impact Study was conducted in 18 sites (comprising 17 school districts and one statewide program) where between 6 and 32 schools per site were assigned to treatment or comparison condition status.[12] Data from the study make it possible to estimate program impacts on RF instructional time (the mediator of interest). In addition, estimates were obtained for program impacts on student reading achievement measured by SAT10 reading scale scores for three annual student cohorts in Grades 1 and 2. The smallest block for estimating impacts is a single cohort in a single grade from a single site. There are 108 such blocks. Because the unit of assignment to RF is school, the effective sample size of these blocks is quite small and the strength of instruments created by interacting assigned treatment status with zero/one block indicators is quite weak (their first-stage *F*-statistic is 3.48). Thus, our analyses are based on 36 blocks (which pool student cohorts within grade-by-site cells) or 18 blocks (which pool student cohorts and grades within sites).

As we reported in the introduction section, an IV analysis with a single instrument indicates that on average, student reading achievement increased by 0.37 scale score points $\left(\frac{4.29}{11.6}\right)$ per additional minute of RF instruction. The right side of Table 3 reports corresponding results obtained from OLS, 2SLS with multiple instruments, and the two bias-corrected IV estimators with multiple instruments. The OLS estimates indicate a very small mediator effect: an additional 0.037 or 0.122 scale score points per minute of RF instruction per daily reading block (which correspond to effect sizes of 0.001 and 0.003 per minute of instruction for 18 blocks or 36 blocks, respectively). The 2SLS estimates of this mediator effect are much larger: 0.397 or 0.387 for 18 or 36 blocks (effect sizes of roughly 0.01), respectively, estimates that are very close to the single instrument estimate of 0.37 points per minute of RF instruction.

The corresponding bias-corrected IV estimates are 0.365 for 18 blocks and 0.484 for 36 blocks. Hence, they are roughly comparable to estimates produced by 2SLS. This is especially true for the finding based on 18 blocks where the first-stage *F*-statistic for 2SLS (17.7) suggests that one can have some confidence in the bias-corrected estimators. This suggests that the RF example might not involve substantial compliance covariance bias. To explore this issue, it would be useful to examine the quadratic coefficient in the regressions used to produce bias-corrected estimates. However, as can be seen from Table 3, this coefficient is not estimated precisely enough to provide information that is useful for this purpose.

Several further points about these findings are important to consider. Note first that estimated standard errors are not presented for the OLS, 2SLS, or

bias-corrected estimators. This is because the small number of schools in each block (the smallest blocks have only six schools) does not support valid bootstrapped standard errors (Freedman, 2005). Thus, for this example, it is not possible to use bootstrapped standard errors to provide statistical inferences for any of the estimators.[13] This problem is likely to arise frequently when aggregate units (clusters) are assigned to treatment or control status, which typically results in small numbers of aggregate units per block. Note second that estimates of mediator effects produced by 2SLS and the bias-corrected estimator are many times larger than those produced by OLS. This probably reflects attenuation bias in the OLS estimates that is created by a lack of reliability in the observational measure of RF instructional time (each classroom was only observed by a single rater during a single 60- to 90-min reading block). Neither 2SLS nor the bias-corrected estimators are subject to this problem.

In summary, Project STAR illustrates a situation in which the bias-corrected estimators are likely to work quite well: The *F*-statistic is unusually large (over 1,000), the coefficient of variation for compliance equals about 0.26, and the number of observations per block (over 70) is large enough to support accurate bootstrapped standard errors. RF provides a much more limited application. The *F*-statistic is 17.7 or 8.2, the coefficient of variation for compliance is 0.76 or 0.79, and the number of observations per block (ranging from 6 to 32) is too small to support bootstrapped standard errors.

VII. Discussion and Conclusion

The use of multiple site-by-treatment status instruments to identify the effects of the mediators of a treatment in a multisite trial is a potentially promising method, though it does not come without some complexity. In addition to the usual set of assumptions required for identification in IV models, an additional assumption—that there is no correlation between the site-average compliance rates and the site-average effects of the mediator—is required (Reardon & Raudenbush, in press). This assumption is required regardless of whether the goal is to identify a complier average causal effect (a LATE, in Angrist et al.'s [1996] 1996 terminology) or an average effect in a population. Note that in 2SLS estimation (and in other parametric IV methods), the assumption of compliance-effect independence implies that the relationship between the site-specific intent-to-treat effects (the β_s's in our notation) and the site-specific average compliances (the γ_s's) is linear. If the compliance-effect independence assumption is not met, then the implicit linearity assumption in multiple-instrument 2SLS model will lead to biased estimates. However, this bias is not unique to 2SLS multiple-site, multiple-mediator-IV estimation: It is present as well in some other standard methods of IV analysis, such as the limited information maximum likelihood IV estimator. Moreover, Raudenbush et al. (2012) note that even a nonparametric method such as averaging estimates from multiple sites (which might themselves be

estimated using any one of a number of IV estimators) using precision weights will suffer from the same between-site compliance-effect covariance bias that we describe here.

Reardon and Raudenbush (in press, online Appendix B) derive an asymptotic expression for the 2SLS bias due to compliance-effect covariance, but do not consider how compliance-effect covariance bias may interact with FSB. Here we have shown that the magnitude of the compliance-effect covariance bias depends on the strength of the instruments. We have derived an analytic expression approximating the magnitude of both FSB and compliance-effect covariance bias. This expression shows that, *ceteris paribus*, the magnitude of compliance-effect covariance bias increases asymptotically as the instruments grow stronger, while FSB decreases. Thus, a strong set of instruments is no guarantee against compliance-effect covariance bias. Our simulations illustrate that the bias formula closely matches the true bias over a wide range of the parameter space and demonstrates that the bias due to compliance-effect covariance may be substantial.

To address this problem, we develop two closely related alternative IV estimators —the bias-corrected IV estimator and the plug-in bias-corrected IV estimator. Our simulations show that these two estimators perform very well over a wide range of conditions when the first-stage *F*-statistic is greater than 10. In this situation, as long as CV_γ is not too extreme and $Corr(\gamma_s, \delta_s)$ is not very close to zero, the bias-corrected estimators generally outperform the 2SLS estimator both in terms of bias and RMSE. Note that both the coefficient of variation for compliance and the first-stage *F*-statistic can easily be estimated based on the data, so researchers can readily assess whether it is preferable to use the bias-corrected estimators.

The two bias-corrected estimators rely on a weaker assumption than the 2SLS estimator. While 2SLS requires the assumption that the site-average compliances and the site-average effects of the mediator are independent, the bias-corrected estimators require only that the association between the site-average compliances and the site-average effects be linear. This is a significantly more plausible assumption than the assumption of no association. The bias-corrected estimators are therefore preferable to 2SLS in a wide range of situations for the analysis of mediator effects in multisite trials.

Several general caveats are important to note here. First, because IV models rely heavily on the exclusion restriction for identification of mediator effects, IV analysis is suitable for mediation analysis only when the exclusion restriction is valid—that is, only when the effect of the instrument on the outcome is fully mediated by the specified mediator or mediators. Partial mediation models, in which there may be a direct effect of the instrument as well as mediated effects, rely on fundamentally different assumptions and different analytic strategies than those we have described in this article. IV models for mediation require that we specify and measure all mechanisms through which an instrument affects an outcome.

Second, our focus in this article has been on reducing the 2SLS bias caused by *between-site* compliance-effect covariance. If, however, the mediator is not binary, and the researcher wishes to estimate an average effect of the mediator in the population, there may be additional bias caused by *within-site* compliance-effect covariance (Reardon & Raudenbush, in press). Such potential bias is a feature of all 2SLS estimators, whether they rely on a single instrument or multiple instruments. In principle, this bias may be larger or smaller than the bias due to between-site compliance-effect covariance, depending on the magnitudes of the covariances and the strength of the instruments. Methods of detecting and correcting such within-site compliance-effect covariance bias have been suggested elsewhere (Heckman & Vytlacil, 1999; Reardon & Raudenbush, in press); we do not discuss them here.

Appendix

Simulation Set-up

The data used in the simulations presented in this paper are generated through a two-step process. In the first step, we generate a set of 50 sites, each characterized by the vector $[\gamma_s, \delta_s, \Lambda_s, \Theta_s, n_s, p_s]'$, drawn from a population where

$$\begin{bmatrix} \gamma_s \\ \delta_s \\ \Lambda_s \\ \Theta_s \\ n_s \\ p_s \end{bmatrix} \sim N\left[\begin{pmatrix} \gamma \\ 1 \\ 0 \\ 0 \\ 200 \\ 0.5 \end{pmatrix}, \begin{pmatrix} \tau_\gamma^2 & \tau_{\gamma\delta} & 0 & 0 & 0 & 0 \\ \tau_{\gamma\delta} & \tau_\delta^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \right] \tag{A.1}$$

We fix $n_s = n = 200$ and $p_s = p = 0.5$ for all simulations here for simplicity, and set the covariances of the site fixed effects in the first and second stage equations (Λ_s and Θ_s in our notation) with every other parameter to be zero. The means of Λ_s and Θ_s are arbitrarily set to 0 and their variances are arbitrarily set to 1, but these means and variances have no impact on the bias or precision of any of the estimators discussed here. By manipulating γ, τ_γ^2, τ_δ^2, and $\tau_{\gamma\delta}$, we can set $CV_\gamma, F, Corr(\gamma_s, \delta_s)$, and $sd(\delta)$, to the values used in Tables 1 and 2. Specifically, we set

$$\begin{aligned} \gamma &= \left(\frac{\sigma^2}{np(1-p)} \cdot \frac{(F-1)}{1+CV_\gamma} \right)^{\frac{1}{2}} = \left(\frac{0.02 \cdot (F-1)}{1+CV_\gamma} \right)^{\frac{1}{2}} \\ \tau_\gamma^2 &= \gamma^2 \cdot CV_\gamma^2 \\ sd(\delta) &= \sqrt{\tau_\delta^2} \\ \tau_{\gamma\delta} &= \left(\tau_\gamma^2 \cdot \tau_\delta^2 \right)^{\frac{1}{2}} \cdot Corr(\gamma_s, \delta_s). \end{aligned} \tag{A.2}$$

These values ensure that the simulations correspond to the scenarios described in Tables 1 and 2.[14]

In the second step, we generate 200 observations within each site, each characterized by the vector $[\Gamma, \Delta, e_i, u_i]'$. The sample in a site s is drawn from a population where

$$\begin{bmatrix} \Gamma \\ \Delta \\ e_i \\ u_i \end{bmatrix} \sim N \left[\begin{pmatrix} \gamma_s \\ \delta_s \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & \rho\sigma\omega \\ 0 & 0 & \rho\sigma\omega & \omega^2 \end{pmatrix} \right]. \tag{A.3}$$

For simplicity, we fix $\sigma^2 = Var(e_i) = \omega^2 = Var(u_i) = 1$ and $\rho = 0.5$ in all simulations. We also set $Var_s(\Gamma) = Var_s(\Delta) = 0$ in all sites. Note that this simulation design constrains compliance and effect to vary (and covary) only across sites; there is no variance among individuals within a site.

We then randomly assign 100 observations within each site to receive $T_i = 1$, and the other 100 to receive $T_i = 0$. We then compute, for each observation, values of the mediator and the outcome:

$$\begin{aligned} M_{is} &= \Lambda_s + \Gamma T_i + e_i \\ Y_{is} &= \Theta_s + \Delta M_{is} + u_{is}. \end{aligned} \tag{A.4}$$

For each simulation scenario, we repeat this process 2000 times to generate the estimates shown in Tables 1 and 2.

Acknowledgments

The authors thank Steve Raudenbush for his invaluable insights and Takoko Nomi and Michael Seltzer for their helpful comments.

Authors' Note

The ideas presented and positions taken in the article are solely the responsibility of the authors, and do not necessarily reflect views of the funders.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the Institute of Education Sciences (R305D090009).

Supplementary Materials

The online appendices are available at http://jeb.sagepub.com/supplemental.

Notes

1. The econometrics literature on instrumental variables analysis of correlated random coefficient models (Heckman & Vytlacil, 1998) addresses an issue that differs somewhat from compliance-effect covariance bias. Bias in correlated random-coefficients models is produced by a correlation between the *level* of a mediator and its per unit effect on an outcome of interest. This would occur, for example, if sites that used more of a particular type of reading instruction experienced larger (smaller) effects on student reading achievement per unit of the instruction than did sites that used less of the instruction. Compliance-effect covariance bias is produced by a correlation between a treatment-induced *change* in the value of a mediator and its per unit effect on an outcome of interest. This would occur, for example, if sites where treatment increased the specific type of reading instruction by a lot experienced larger (smaller) effects per unit of the instruction on student achievement than did sites where treatment increased the instruction by less.
2. On the other hand, if the extra instruments explain little additional variance in the mediator, using K site-by-treatment assignment instruments may produce multiple weak instruments, leading to inefficient and biased estimates (Chamberlain & Imbens, 2004; Staiger & Stock, 1997). A fourth possible approach is to use J instruments, where $1 < J < K$, by interacting treatment status with indicators for J subsets of the K sites, where the subsets are defined in such a way that there is little within-subset variation in compliance. We take up this possibility later in the article.
3. To see this, note that $F = \frac{np(1-p)}{\sigma^2}\left(\gamma^2 + \tau_\gamma\right) + 1$, so we can write the compliance-effect bias term as $2\frac{np(1-p)}{\sigma^2}\gamma \mathrm{Cov}(\gamma_s, \delta_s)\left(\frac{1}{F}\right)$, so $\gamma = 0$ implies the bias is zero.
4. Note that if $\gamma \neq 0$, $F = \frac{np(1-p)}{\sigma^2}\gamma^2\left(1 + \mathrm{CV}_\gamma^2\right) + 1$, that is, F depends on n, p, σ^2, γ, and CV_γ. Therefore, changing F by changing CV_γ will affect compliance-effect covariance bias in two ways, while changes in F due to changes in $n, p, \sigma^2, \mathrm{or}\gamma$, holding CV_γ constant, will only affect compliance-effect covariance bias through their effect on F.
5. Angrist (1990) does this graphically, in a way that is equivalent to weighting each site by the variance of the treatment; in the stylized example here, we assume all sites have equal instrument variance and equal sample size.
6. We can compute F from the estimates of γ, τ_γ, and σ^2 using Equation 3. In practice, if τ_γ is small relative to the sampling variance of the $\hat{\gamma}_s$'s, fitting a random-coefficient model like Equation 12 may not be possible, because the maximum likelihood algorithm may not converge. In such cases, however, there is little or no need to use a random-coefficient model; a fixed effects IV model (a model with a single instrument) would be preferable. We could also fit Equation 12 using site-fixed effects and site-by-treatment assignment

interactions via ordinary least squares (OLS), and then shrink the resulting $\hat{\gamma}_s$'s, as described following Equation 15.

7. In online Appendix A5, we discuss the case where F is small and/or CV_γ is small; in such cases, the final term in Equation 15 may have a large, nonzero expected value, implying that Equation 15 should have a nonzero intercept. In such cases, however, our simulations show that including an intercept in model (Equation 15) leads to a very large sampling variance of the estimates of the intercept and $\hat{\alpha}_0$; the loss in precision is far worse than any reduction in bias achieved.
8. Some iterations did not produce an estimate for δ because the restricted maximum likelihood model used to obtain shrunken estimates of γ_s did not converge. Therefore, the actual number of successful iterations varies by parameter values used in the simulation, ranging from 1,821 to 2,000 of the 2,000 total iterations.
9. Panel A demonstrates this pattern for an F-statistic of 10. Additional results (not reported here) demonstrate that while this pattern holds for a wide range of F-statistics, this "U" shape pattern is more pronounced when the F-statistic is small and becomes more muted as the F-statistic increases.
10. This figure shows how the four estimators behave as CV_γ starts to deviate from the optimal value of 1 toward 0. Results are similar to those presented here when CV_γ deviates from the optimal value of 1 toward infinity. Figure B1 in online Appendix B present graphical demonstrations of those results.
11. Like in Figure 2, this figure presents situations when CV_γ deviates from 1 toward 0. Results are similar when CV_γ deviates from 1 toward infinity. Results for those cases are presented in Figure B2 of online Appendix B.
12. Treatment was not assigned randomly in most of the Reading First (RF) sites, but was rather assigned on the basis of an observed rating score. Our analysis here, like the impact analysis reported by Gamse, Bloom, Kemple, and Jacob (2008), is based on a regression discontinuity design, but that feature of the analysis is not essential to our exposition and so is excluded for simplicity.
13. For the two-stage least squares (2SLS) and ordinary least squares (OLS) estimators, it is possible to obtain estimated standard errors through conventional methods based on standard software packages. However, as demonstrated earlier in the article, those standard errors tend to understate the sampling variation, especially when first-stage F is small. Therefore, conventional standard errors for the OLS and 2SLS estimators are not reported in Table 3 either.
14. The 0.02 term in Equation A.2 comes from the fact that we set $\sigma^2 = 1$ below.

References

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, *80*, 313–336.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Angrist, J., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, *77*, 531–553.

Bound, J., Jaeger, A., & Baker, R. (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.

Chamberlain, G., & Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica*, *72*, 295–306.

Duncan, G. J., Morris, P., & Rodrigues, C. (2011). Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment Experiments. *Developmental Psychology*, *47*, 1263–1279.

Finn, J., & Achilles, C. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*, 557–577.

Freedman, D. A. (2005). *Statistical models: Theory and practice*. New York, NY: Cambridge University Press.

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). *Reading first impact study: Interim report* (NCEE 2008-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica*, *70*, 163–189.

Heckman, J. J., & Vytlacil, E. (1998). Instrumental variable methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *The Journal of Human Resources*, *33*, 974–987.

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, *98*, 4730–4734.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*, 83–119.

Krueger, A. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, *114*, 497–532.

Nomi, T., & Raudenbush, S. (2012). *The impact of math curricular reform on course-taking, classroom composition and achievement: A multi-site discontinuity design* (Working Paper). Paper presented at the Fall conference of the Association for Public Policy Management and Analysis, Baltimore, MD.

Nye, B., Hedges, L., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, *37*, 123–151.

Raudenbush, S., Reardon, S., & Nomi, T. (2012). Statistical analysis for multi-site trials using instrumental variables. *Journal or Research on Educational Effectiveness*, *5*, 303–332.

Reardon, S., & Raudenbush, S. (in press). Under what assumptions do site by treatment instruments identify average causal effects? *Sociological Methods and Research*.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers (New Series)*, *3*, 135–146.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Staiger, D., & Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*, 557–586.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Donald W. K. Andrews & James H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge, England: Cambridge University Press.

Authors

SEAN F. REARDON is a professor of education and (by courtesy) sociology at Stanford University. He is a specialist in methods of measuring social and educational inequality (including the measurement of segregation and achievement gaps) and methods of causal inference in educational and social science research. He conducts research on the effects of educational policy on educational and social inequality, on the causes, patterns, trends, and consequences of social and educational inequality, and in applied statistical methods for educational research.

FATIH UNLU is a senior associate/scientist at Abt Associates, Cambridge, MA. His primary research interests are research methods for education policy evaluations and economics of education.

PEI ZHU is an economist specializing in education policy at MDRC, a nonprofit social and educational policy research organization located in New York, NY. Her current work focuses on experimental and quasi-experimental impact analyses, evaluation design, and related methodological issues. She has been playing a leading role in the student achievement impact analysis for several large-scale randomized experiments and quasi-experiments. Her work at MDRC also includes several methodological studies on empirical issues related to quasi-experimental designs, group randomized experiments, and reliability of measurements for group settings.

HOWARD S. BLOOM is chief social scientist of MDRC, a nonprofit social and educational policy research organization located in New York, NY. He is a specialist in evaluation research methodology with a focus on improving and using experimental and quasi-experimental designs. He has published widely and given many workshops on these issues to applied social scientists in a wide range of fields. In addition, he has designed and led the analysis of numerous large-scale evaluations of programs in the fields of education, employment, welfare, housing, and correctional policy.

Manuscript received August 29, 2012
Revision received April 20, 2013
Accepted September 27, 2013