# Bias and Bias Correction in Multi-Site Instrumental Variables Analysis Of Heterogeneous Mediator Effects

Sean F. Reardon, *Stanford University*
Fatih Unlu, *Abt Associates*
Pei Zhu, *MDRC*
Howard Bloom, *MDRC*

Version: August 29, 2012

This work was supported by a grant from the Institute for Education Sciences (R305D090009). The authors thank Steve Raudenbush for his invaluable insights and Takoko Nomi and Michael Seltzer for their helpful comments. The ideas presented and positions taken in the paper are solely the responsibility of the authors however, and do not necessarily reflect views of the funders.

**Bias and Bias Correction in Multi-Site Instrumental Variables Analysis**

**Of Heterogeneous Mediator Effects**


## Abstract

We explore the use of instrumental variables (IV) analysis with a multi-site randomized trial to estimate the effect of a mediating variable on an outcome. 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 treatment effect) to vary across sites and to co-vary with one another. This extension of conventional fixed-coefficient IV analysis illuminates a potential bias in IV analysis which Reardon and Raudenbush (forthcoming) 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 co-varying) compliance and treatment effects. We next develop an alternate IV estimator that is less susceptible to compliance-effect covariance bias. We compare the bias, sampling variance, and root mean squared error of this "bias-corrected IV estimator" to those of 2SLS and OLS. We find that, when the first stage $F$-statistic exceeds 10 (a commonly-used threshold for instrument strength), the bias-corrected estimator typically performs better than 2SLS or OLS. In the last part of the paper we use both the new estimator and 2SLS to reanalyze data from two large multi-site studies.

**Bias and Bias Correction in Multi-Site Instrumental Variables Analysis**

**Of Heterogeneous Mediator Effects**

## 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 policy making and practice. However, because the theories behind most interventions are not well-developed, and because many interventions have multiple components, it has not yet been possible to determine the causal factors that produce or "mediate" the intervention effects that have been observed. In other words, it has not yet been possible to document the "active ingredients" of these interventions.

Even in a randomized experiment, ordinary least squares regression cannot be relied on to provide unbiased estimates of the effect of a mediating variable 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).[1] Insrtumental variables (IV) methods, however, 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

[1] The econometrics literature refers to this problem as "errors-in-variables" (Greene, 1993).

can be assumed that the full effect of the treatment on the outcome is produced by the mediator, the average effect of the mediator on the outcome ($\frac{\Delta Y}{\Delta M}$) equals the ratio of the effect of the treatment on the outcome ($\frac{\Delta Y}{\Delta T}$) to the effect of the treatment on the mediator ($\frac{\Delta M}{\Delta T}$). Because the randomized experiment or regression discontinuity design provides unbiased estimates of the latter two effects, their ratio will be an (approximately) unbiased estimate of the effect of a unit change in the mediator on the outcome ($\frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{\Delta Y}{\Delta T}$).[2]

Consider for example, the recent multi-site impact evaluation of the federal Reading First (RF) Program (Gamse et. al., 2008) on reading achievement in the early elementary school grades. Reading First'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. The results of the RF impact study showed that on average, Reading First increased the amount of time that teachers spent on RF instruction by 11.6 minutes per day ($\frac{\Delta M}{\Delta T} = 11.6$) and increased student reading achievement by 4.29 scale score points ($\frac{\Delta Y}{\Delta T} = 4.29$). If all of Reading First'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 ($\frac{\Delta Y/\Delta T}{\Delta M/\Delta T} = \frac{4.29}{11.6} = 0.37$).

The Reading First study was a multi-site trial, in which schools in 18 states were assigned by randomization or on the basis of a continuous rating score to receive the RF program or not. In a multi-site design, a more complex IV analysis is possible. Because the treatment is ignorably

[2] Strictly speaking, the ratio will only be unbiased in infinite samples, because estimation error in the numerator and denominator will cause "finite sample bias" in the ratio (Bound, Jaeger, and Baker, 1995). However, if the estimation error in $\Delta M/\Delta T$ is small, this bias will likewise be small.

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 Reading First increased the use of RF instruction by 20 minutes per daily reading block in some sites and by 2 minutes 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 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 (forthcoming), and Nomi and Raudenbush (forthcoming).

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 (Bound, Jaeger, & Baker, 1995; Hahn & Hausman, 2002; Stock & Yogo, 2005; and Angrist & Pischke, 2009).

In this paper we investigate the magnitude of the bias of multiple-site, multiple instrument instrumental variables estimators. We consider not only the role of finite sample bias, but also the role of a second type of bias, what Reardon and Raudenbush (forthcoming) 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 both covary across sites (or persons, though in the present paper

we are concerned with between-site variation).[3] Reardon and Raudenbush (forthcoming) 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 paper 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 finite sample bias, 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 paper, we develop a "bias-corrected IV estimator" that is designed to reduce bias caused by compliance-effect covariance across sites. We use simulations to compare the statistical properties of this new estimator to those of 2SLS and OLS. These findings indicate that under a wide range of conditions, the new estimator performs better than 2SLS and OLS (in terms of bias and mean squared error) if the instruments used have a first-stage $F$-statistic greater than 10.

The paper concludes with two examples of the application of the bias-corrected IV estimator. We first use it to estimate the effect of class size on student achievement, using data from data for the Tennessee class-size experiment, Project STAR. We then use it to reanalyze data

[3] 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.

from the Reading First Impact Study described above, 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 multi-site randomized trial, in which $N$ subjects (indexed by $i$) are nested in a set of $K$ sites (indexed by $s \in \{1,2,\dots,K\}$). Within each site, a random sample of $n = N/K$ subjects 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 $\Gamma$) and the person-specific effect of $M$ on $Y$ (the person-specific "effect," denoted $\Delta$) may be heterogenous across subjects. Our goal is to estimate the average effect of $M$ on $Y$ in the population, denoted $\delta = E[\Delta]$.

We will assume that $cov_s(\Gamma,\Delta) = [cov(\Gamma,\Delta)|S = s] = 0$ (no within-site compliance-effect covariance). This assumption is met unambiguously if both $T$ and $M$ are binary and we focus only on compliers (Reardon & Raudenbush, forthcoming). However, we will not assume that the between-site compliance-effect covariance (denoted $cov(\gamma_s,\delta_s)$, where the average compliance in a site $s$ is denoted $\gamma_s$ and the average effect of $M$ on $Y$ in site $s$ is denoted $\delta_s$) is zero.

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

$$M_i = \Lambda_s + \gamma_s T_i + e_i, \quad e_i \sim N(0,\sigma^2)$$

$$Y_i = \Theta_s + \delta_s M_i + u_i, \quad u_i \sim N(0,\omega^2)$$

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\omega \\ \rho\sigma\omega & \omega^2 \end{pmatrix} \right],$$

where $\rho$ is the correlation between $e$ and $u$. Across sites, the covariance matrix of the $\gamma_s$'s and the $\delta_s$'s is

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

*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 instrumental variables methods, and then to average the $\delta_s$'s across sites. There are several drawbacks to this approach, however. First, if the instrument is weak in some sites, the estimated $\delta_s$ in those sites may be substantially biased due to finite sample bias, leading to bias in the estimated average effect. Second, a precision-weighted average of the $\delta_s$ will weight sites with greater compliance (larger values of $\gamma_s$) more, leading to biased estimates of $\delta$ 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, Reardon, & Nomi, 2012). This model would assume, implicitly, that $\gamma_s$ is constant across sites. If $\gamma_s$ 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 over-identified IV model, using $K$ site-by-treatment status interactions as instruments. In some cases, such a model may be preferable to either of the two cases above. 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. 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, as is done by Kling, Liebman, and Katz (2007), Duncan, Morris, and Rodrigues (forthcoming), and Nomi and Raudenbush (2012) (this approach is discussed in Reardon & Raudenbush, forthcoming). Although we do not consider the multiple mediator case in this paper, 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

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

The second stage equation is

$$Y_i = \Theta_s + \delta M_i + u_i, u_i \sim N(0, \omega^2). \tag{2b}$$

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

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

(4a)

Estimating $\delta$ via two-stage least squares (2SLS) will also result in bias. In particular, as we show in Appendix A3, the 2SLS bias (the bias in the estimate of $\delta$ obtained from fitting Equations (2a) and (2b) via 2SLS) is approximately

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

(5a)

Note that both the OLS bias and the 2SLS bias have two components— one component that depends on the covariance of the errors ($\rho$), and one component that depends on the covariance between the gammas and deltas ($\tau_{\gamma\delta}$). The first component can be thought of as bias that arises from *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 finite sample bias in IV estimators. The second component can be thought of as bias that arises from *selection on effects* (individuals' compliance with the instrument is correlated with the effect the mediator has on their value of the outcome $Y$), as would be predicted by the Roy model (Roy, 1951; Borjas, 1987); it gives rise to what we refer to as compliance-effect covariance bias (Reardon & Raudenbush, forthcoming). 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 2SLS biases can be written as weighted averages of the two components:

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

(4b)

and

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

(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 $n \gg F$, $\lambda^{OLS}$ 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 (finite sample bias) is minimized and bias due to the correlation of $\gamma_s$ and $\delta_s$ 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; conversely, the bias due to the first component will always get more weight in the OLS estimator than in the 2SLS 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 finite sample bias. 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, finite sample bias becomes trivial.

The second component of the bias in (5a) is compliance-effect covariance bias. If $\gamma = 0$, this bias term is 0.[4] If, however, $\gamma \neq 0$, we can write the compliance-effect covariance bias term as

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

where $CV_\gamma = \sqrt{\tau_\gamma}/\gamma$ is the coefficient of variation of $\gamma_s$.

The compliance-effect covariance bias component depends on four factors. First, the bias term is linear in the correlation between $\gamma_s$ and $\delta_s$. Stronger correlations produce larger biases.

[4] 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 $2\frac{np(1-p)}{\sigma^2}\gamma Cov(\gamma_s, \delta_s)\left(\frac{1}{F}\right)$, so $\gamma = 0$ implies the bias is zero.

Second, the bias term is linear in the standard deviation of the $\delta_s$'s across sites. Greater between-site variation in the effects of the mediator leads to greater bias. 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 $Corr(\gamma_s, \delta_s), \tau_\delta$, and $F$, the magnitude of the compliance-effect covariance is maximized when $|CV_\gamma| = 1$ (see appendix A4). As $CV_\gamma$ 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 is smaller when $F$ is small. When the instruments are collectively strong, the bias due to between-site compliance-effect covariance is maximized.[5]

The two bias components—finite sample bias and compliance-effect covariance bias—are oppositely affected by $F$. When $F$ is large, finite sample bias is trivial, but compliance-effect covariance is maximized. That is, compliance-effect covariance can lead to bias in the 2SLS estimator even with an arbitrarily strong set of instruments (i.e., with arbitrarily large within-site samples). When $F$ is small, compliance-effect covariance bias is reduced, but finite sample bias is maximized.

Each of the four factors influencing the compliance-effect covariance bias component is, in principle, estimable from the observed data (although estimation of $\tau_\delta$ and $Corr(\gamma_s, \delta_s)$ will be complicated by finite sample bias in the estimation of the $\delta_s$'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 finite sample bias to the overall bias is negligible. This suggests that we may be able to devise a better estimator of $\delta$—one that is less biased by compliance-effect covariance—than 2SLS, at least for the case where $F$ is relatively large. In Part IV of this paper, we

[5] Note that if $\gamma \neq 0, F = \frac{np(1-p)}{\sigma^2}\gamma^2\left(1 + CV_\gamma^2\right) + 1$, i.e., $F$ depends on $n, p, \sigma^2, \gamma$, and $CV_\gamma$. Therefore, changing $F$ by changing $CV_\gamma$ will affect compliance-effect covariance bias in two ways while changes in $F$ due to changes in $n, p, \sigma^2$, or $\gamma$, holding $CV_\gamma$ constant, will only affect compliance-effect covariance bias through their effect on $F$.

develop such an estimator.

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 $\delta_s$ is constant across sites (Chamberlein and Imbens, 2004; Angrist and Pischke, 2009). 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: (i) 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, (ii) to assess the sampling variation of the 2SLS estimator in the presence of compliance and effect heterogeneity and compliance-effect covariance; and (iii) to compare the magnitude of the bias and the root mean squared error (RMSE) of the 2SLS estimator relative to the OLS estimator. All simulated data represent the case with 50 sites and 200 individuals per site, half of whom are randomized to treatment and half of whom are randomized to control status. 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 $\gamma_s$ and $\delta_s$ *across-site*, not within-site. Appendix B provides a more detailed description of the simulation set-up.

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 — $CV_\gamma$, the

expected first-stage *F*- statistic, the compliance-effect correlation, and the variance of the effect, $\tau_\delta$— is systematically manipulated while the other three parameters are held constant (see Appendix B for details). Panel A varies $CV_\gamma$ between 0 and infinity (column 1 in Table 1) while keeping the target *F*-statistic at 10, a commonly used threshold for reasonably strong instruments (Staiger and Stock, 1997; Stock and Yogo, 2003), which is achieved by manipulating the compliance mean and standard deviation to yield the desired $CV_\gamma$ and *F*-statistic. Panel B in Table 1 varies the expected $F$-statistic between 2 and 101 (column 2). Panel C varies the compliance-effect correlation between -0.75 and 0.75, including a scenario where the correlation is set to zero (column 3). Finally, Panel D varies $\sqrt{\tau_\delta}$ between 0 and 5 (column 4) while holding $\delta$ fixed at 1. The rest of the columns report the results obtained from 2000 simulation samples drawn from a population of sites generated according to the parameter values shown in columns 1-4. We describe below the main results of these simulations.

*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 2000 simulations and the true effect). In each case, the estimated bias is very close to that predicted by Equation (5a). As expected, Table 1 shows that the bias is larger when $CV_\gamma$ is near 1; when $F$ is small; when the correlation of $\gamma_s$ and $\delta_s$ is large; and when the variance of $\delta_s$ 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 $\delta$ is large (see rows 11, 15, and 19).[6]

[6] Note that the simulated data is generated based on a true effect of $\delta = 1$, so the bias reported in columns 5 and 6 can be interpreted the ratio of the bias to the magnitude of the true effect.

*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 $\delta$ across the 2,000 simulation samples) and the average standard error reported by conventional 2SLS estimation algorithms (column 8).[7] These conventional 2SLS standard errors are based on the assumption that $\delta_s$ is constant across sites. Equation (B14) in Reardon and Raudenbush (forthcoming), however, implies that the sampling variance of the 2SLS estimator depends on the variance of $\delta_s$; 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 standard errors. Only in row 16, where $\tau_\delta$ 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 over-identified 2SLS models, the estimated standard errors are often too small, especially when the instruments are collectively weak (Chamberlein and Imbens, 2004; Angrist and Pischke, 2009). Even in row 10, where the $F$-statistic is 101, the 2SLS standard error is one-tenth of the true standard error. We conclude that conventional 2SLS 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

[7] We use Stata version 12 (Statacorp, 2011) for these analyses.

deviation of the OLS estimates across the 2,000 simulation samples in each case); column 13 reports the average reported OLS standard error across the 2000 samples; and column 14 shows the root mean squared error (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 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 standard errors substantially underestimate sampling variance when the mediator effect is heterogenous 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. **A Bias-Corrected Multi-Site Single Mediator IV Estimator**

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 bias-corrected estimator, consider the hypothetical data described in Figure 1 below. Each of the panels on the left side of the figure shows a the relationship between $\delta_s$ and $\gamma_s$. In each case, $\delta$ (the average value of $\delta_s$ across sites) equals 1. Likewise, in each case, the average compliance across sites equals 1, and both $\gamma_s$ and $\delta_s$ 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 $F$, $Corr(\gamma_s, \delta_s)$, and $\tau_\delta$). The three figures on the left side differ only in the correlation between $\delta_s$ and $\gamma_s$, ranging from $Corr(\delta_s, \gamma_s) = -0.50$ to $Corr(\delta_s, \gamma_s) = +0.50$.

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 ITT effects against the $\gamma_s$'s. In practice, we can estimate the $\beta_s$'s and the $\gamma_s$'s, so we can readily produce figures of the type shown here. Note that a non-zero correlation between $\gamma_s$ and $\delta_s$ will produce a figure on the right that shows a non-linear association between $\beta_s$ and $\gamma_s$. This is evident in the quadratic fitted curves in the righthand figures. Thus, non-linearity in the observed relationship between $\beta_s$ and $\gamma_s$ 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 $\beta_s$ on $\gamma_s$ (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 (Reardon & Raudenbush, forthcoming; Raudenbush, Reardon, & Nomi, 2012).[8] The slope of this line is the 2SLS IV estimate of $\delta$. Recall that the average

[8] 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.

value of $\delta_s$ 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 in the figures. Note that when $Corr(\delta_s, \gamma_s) > 0$, the slope of the fitted line is substantially greater than 1, and when $Corr(\delta_s, \gamma_s) < 0$, the slope of the line is substantially less than 1. The reason for this is that the sites where $\gamma_s$ is largest (farthest from 0) have more leverage in the regression; the non-zero correlation between $\gamma_s$ and $\delta_s$ means that these sites also have larger than average $\delta_s$'s, which results in bias in the estimates.

*The Bias-Corrected Estimator*

We now develop a bias-corrected IV estimator. First, assume that the association between $\gamma_s$ and $\delta_s$ is linear:[9]

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

Taking the expectation of both sides of Equation (7) yields

$$E[\delta_s] = \alpha_0 + \alpha_1 E[\gamma_s]$$

$$\delta = \alpha_0 + \alpha_1\gamma. \tag{8}$$

This suggests that if we could estimate $\alpha_0$, $\alpha_1$, and $\gamma$, we can estimate $\delta$ as

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

First, we can estimate $\gamma$ from the following random-coefficients model:[10]

[9] Note that this assumption is weaker than the assumption that $Cov(\gamma_s, \delta_s) = 0$. We can, in principle, relax the linearity assumption further, and allow the relationship between $\gamma_s$ and $\delta_s$ 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 $\gamma_s$. This would result in a higher-order regression model in Equation (12) below.

[10] In practice, if $\tau_\gamma$ is small relative to the sampling variance of the $\hat{\gamma}_s$'s, fitting a random coefficient model like (10) 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

$$M_i = \Lambda_s + \gamma_s T_i + e_i$$

$$\begin{pmatrix} \Lambda_s \\ \gamma_s \end{pmatrix} \sim N\left[\begin{pmatrix} \Lambda \\ \gamma \end{pmatrix}, \begin{pmatrix} \tau_\Lambda & \tau_{\gamma\Lambda} \\ \tau_{\gamma\Lambda} & \tau_\gamma \end{pmatrix}\right].$$

(10)

Now, we note that

$$\beta_s = E[B|S = s]$$

$$= E[\Gamma\Delta|S = s]$$

$$= E[\Gamma|S = s] \cdot E[\Delta|S = s] + Cov(\Gamma\Delta)|S = s$$

$$= \gamma_s \cdot \delta_s + Cov(\Gamma\Delta)|S = s.$$

(11)

Given the assumption of no within-site compliance-effect covariance, substituting Equation (7) into (11) yields

$$\begin{aligned} \beta_s &= \gamma_s \cdot \delta_s \\ &= \gamma_s(\alpha_0 + \alpha_1\gamma_s + \nu_s) \\ &= \alpha_0\gamma_s + \alpha_1\gamma_s^2 + \gamma_s\nu_s, \;\; \nu_s \sim N[0, \sigma_\nu^2]. \end{aligned}$$

(12)

In other words, under the assumption that $\delta_s$ is linearly related to $\gamma_s$, $\beta_s$ can be written as a quadratic function of $\gamma_s$, passing through the origin, with a heteroskedastic error term. The parameters $\alpha_0$ and $\alpha_1$ 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 $\beta_s$'s and $\gamma_s$'s, two factors will complicate the estimation of $\alpha_0$ and $\alpha_1$ from the observed data. First, we do not observe $\beta_s$ and $\gamma_s$; 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 $\alpha_0$ and $\alpha_1$ because of the error in $\hat{\gamma}$. Second, in finite samples, the correlation between $e$ and $u$ (the

---

single instrument) would be preferable.

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 Appendix A3); this will induce bias in the estimates of $\alpha_0$ and $\alpha_1$.

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

$$E\left[\hat{\beta}_s \middle| \hat{\gamma}_s\right] = \alpha_0 \gamma_s^* + \alpha_1 \gamma_s^{2^*} + Cov(b_s, g_s) \frac{\lambda(\hat{\gamma}_s - \gamma)}{\tau_\gamma}, \tag{13}$$

where $\lambda = \tau_\gamma/(\tau_\gamma + \tau_g)$ is the reliability of the $\hat{\gamma}_s$'s; $\gamma_s^* = E\left[\gamma_s \middle| \hat{\gamma}_s\right] = \lambda\hat{\gamma}_s + (1-\lambda)\gamma$; and $\gamma_s^{2^*} = E\left[\gamma_s^2 \middle| \hat{\gamma}_s\right] = \gamma_s^{*2} + \tau_\gamma(1-\lambda)$. When $F$ is large and $CV_\gamma$ is not small, the expected value of the final term in Equation (13) will be small.[11] This suggests we can regress the $\hat{\beta}_s$'s on $\gamma_s^*$ and $\gamma_s^{2^*}$ (with no intercept) to estimate $\alpha_0$ and $\alpha_1$. Given the estimates $\hat{\gamma}$, $\hat{\alpha}_0$, and $\hat{\alpha}_1$, we can then compute

$$\hat{\delta} = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\gamma}. \tag{14}$$

If we have large samples within sites, we can estimate the $\hat{\beta}_s$'s and $\gamma_s$'s very reliably, which will lead to precise estimates of $\alpha_0$, $\alpha_1$, and $\gamma$, and thus, to precise estimates of $\delta$.

*Standard errors for $\hat{\delta}$*

We compute a standard error for $\hat{\delta}$ via bootstrapping. Specifically, we (i) draw a sample of $K$ sites, with replacement, from the original sample of sites; (ii) draw a sample of $p \cdot n$ treatment and $(1-p)n$ control cases, with replacement, separately in each resampled site; (iii) estimate $\hat{\delta}$ from this new sample as described above in Equation (14); (iv) repeat steps (i)-(iii) many times (we

[11] In Appendix A5 we discuss the case where $F$ is small and/or $CV_\gamma$ is small; in such cases, the final term in (13) may have a large, non-zero expected value, implying that Equation (13) should have a non-zero intercept. In such cases, however, our simulations show that including an intercept in model (13) 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.

use 500 draws in the simulations described below); and (v) use the distribution of the estimates from these repeated draws as an estimate of the sampling variance of $\hat{\delta}$.

## V. **Simulation Analyses**

We assess the performance of the bias-corrected IV estimator described in Section IV using a set of simulations, comparing the results based on the new estimator with those from 2SLS. Appendix B describes the simulation set up in detail. We vary three parameters—the coefficient of variation compliance ($CV_\gamma$), the expected $F$ statistic, and the compliance-effect correlation—across simulations.

Table 2 presents the estimated bias, estimated standard error, and root mean square error of the bias-corrected estimator for a range of simulated populations. In Panel A varies $CV_\gamma$ while holding the expected $F$-statistic constant at 10 and the compliance-effect correlation constant at 0.25. Panel B varies $F$ between 2 and 101, fixing $CV_\gamma = 1$ and $Corr(\gamma, \delta) = 0.25$. Panel C varies the compliance-effect correlation between 0 and 0.75, when $F = 26$ and $CV_\gamma = 1$. Columns 4 through 7 report the estimation results for the bias-corrected estimator. For comparison, Columns 8-11 report the corresponding 2SLS bias, standard error and RMSE.

*Bias of the bias-corrected estimator in the presence of compliance-effect covariance in finite samples*

Column 4 in Table 2 presents the estimated bias for the bias-corrected estimator across 2000 simulation iterations.[12] Panel A indicates that the magnitude of the estimated bias of the bias-corrected estimator reaches its *minimum* value when $CV_\gamma$ equals 1, other things being equal. As $CV_\gamma$

[12] Some iterations did not produce an estimate for $\delta$ because the restricted maximum likelihood (RMLE) model used to obtain shrunken estimates of $\gamma_s$ 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 out of 2,000 total iterations.

deviates from 1, the absolute value of estimated bias increases.[13] Thus, the bias corrected estimator is most effective at eliminating bias when $CV_\gamma$ is near 1. This is in stark contrast with the pattern observed in panel A of Table 1, which shows that bias in 2SLS exhibits an inverse "U" shape that reaches its maximum value when $CV_\gamma$ is 1 and diminishes steadily as $CV_\gamma$ starts deviates from 1 in either direction.

Panel B suggests that the bias-corrected estimator does a good job eliminating bias when the first stage $F$-statistic is large. A comparison between Columns 4 and 8 indicates that, when $F$ is extremely small, the absolute value of bias of the bias-corrected estimator is similar to that of 2SLS. As $F$ increases, the magnitude of the bias in Column 4 decreases both in absolute term and as a proportion of 2SLS bias.

Panel C shows that, for cases examined here, bias in the bias-corrected estimator decreases as $Corr(\gamma_s, \delta_s)$ increases, other things being equal. Note that compliance-effect bias in 2SLS or OLS increases with $Corr(\gamma_s, \delta_s)$. So results in this panel indicate that, when $CV_\gamma$ is 1 and the $F$-statistic is fairly large (e.g., $F = 26$), the bias-corrected estimator performs better when the compliance-effect bias is larger..

*Sampling variability of the bias-corrected estimator in the presence of compliance-effect covariance in finite samples*

Columns 5 and 6 report the true sampling variation (the standard deviation of the estimates across the 2000 simulation samples) and the average bootstrapped standard error for each scenario. In general, except when $F$ is very small, the sampling variance of the bias-corrected estimator 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

[13] 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.

variances of the bias-corrected estimator and of 2SLS are much larger than the conventional 2SLS standard errors, as shown in column 10). Moreover, the bootstrapped standard errors for the bias-corrected estimator are very close to the true standard errors, except when $F$ is very small.

*Comparing the bias-corrected estimator 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 $CV_\gamma$ values of 1.0, 0.2, and 0.[14] In each case, $Corr(\gamma_s, \delta_s)$ 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 grey area on top represents the amount of compliance-effect bias (CEB) in the 2SLS estimator and the dark grey area at the bottom represents the finite sample bias component (FSB) of the 2SLS estimator. This decomposition is based on Equation 5a and the sum of these two components closely tracks the estimated 2SLS bias.[15] These three graphs illustrate that the relative bias of 2SLS and the bias-corrected estimator depends both on $CV_\gamma$ and the first stage $F$-statistic. As expected, the bias corrected estimator reduces 2SLS bias the most when 2SLS compliance-effect bias is large relative to the 2SLS finite sample bias.

Specifically, when $CV_\gamma$ is 1, the bias-corrected estimator always has 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 estimator bias reaches its minimum value when $CV_\gamma$ is 1, while 2SLS bias maximizes at this point. The dotted line closely tracks the FSB area (in dark grey),

[14] This figure shows how the three estimators behave as $CV_\gamma$ starts to deviate from the optimal value of 1 towards 0. Results are similar to those presented here when $CV_\gamma$ deviates from the optimal value of 1 towards infinity. Figure C1 in appendix C present graphical demonstrations of those results.
[15] The sum of these two bias components closely but not exactly tracks estimated bias because the decomposition is based on Equation 5a which is an approximation.

indicating that, in this case, the bias-corrected estimator is very successful in eliminating almost all of the compliance-effect bias in the 2SLS estimator, regardless of the $F$-statistic.

When $CV_\gamma$ is different from 1 but does not lie in the extremes (i.e., $CV_\gamma$=0.2), the bias-corrected estimator can still produce a smaller bias than the 2SLS method if the $F$-statistic is greater than 10 (middle graph). As $CV_\gamma$ continues to deviate from 1 and reaches the extreme of zero (i.e., when $\gamma_s$ does not vary across sites), the bias in the bias-corrected estimator decreases and approaches the bias in the 2SLS estimator as the $F$-statistic increases, but the 2SLS estimator produces the smallest bias among the three methods for all $F$-statistics presented here (bottom graph). This is not surprising since in this case, there is no compliance-effect bias 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 root mean squared error (RMSE) of these three 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 estimator is 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 estimator has the smallest RMSE when $F$ is above some threshold, though this threshold depends on $CV_\gamma$ —it is the smallest when $CV_\gamma$ is 1 (top graph) and becomes larger as $CV_\gamma$ 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 to vary. In these figures, the $F$-statistic is set to a value of 26, and the horizontal axis indicates values of $Corr(\gamma_s, \delta_s)$. All other attributes of the graph are the same as in Figure 2.[16]

[16] Like in Figure 2, this figure presents situations when $CV_\gamma$ deviates from 1 towards 0. Results are similar when $CV_\gamma$ deviates from 1 towards infinity. Results for those cases are presented in Figure C2 of Appendix C.

Similar to Figure 2, the three graphs on the left side of figure 3 show that when $CV_\gamma = 1$, the bias-corrected estimator works well in eliminating the compliance-effect bias in the 2SLS bias, especially when the CEB is large (top graph). When $CV_\gamma$ deviates somewhat from 1, the bias-corrected estimator eliminates some, but not all of the compliance-effect bias (middle graph). When there is no compliance-effect bias in the 2SLS estimator (either because $Corr(\gamma_s, \delta_s) = 0$ 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. 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).

It is clear that the combination of $CV_\gamma$, the $F$-statistic, and $Corr(\gamma_s, \delta_s)$ affect the performance of the bias-corrected IV estimator relative to that of the 2SLS estimator. In general, when the *F*-statistic is greater than 10, the bias-corrected IV estimator outperforms the 2SLS estimator both in terms of bias and RMSE under a wide range of conditions. This is especially true when $CV_\gamma$does not deviate from 1 too much and when $Corr(\gamma_s, \delta_s)$ is not very close to zero. When the $F$-statistic is less than 10, the bias-corrected estimator generally performs worse than 2SLS. However, because IV methods should generally not be used when the $F$-statistc is less than 10 (for example, see Yogo and Stock, 2005), this is not a particularly useful comparison.

## VI. **Empirical Examples**

We now apply 2SLS and the bias-corrected IV estimator to the reanalysis of data from two studies: (1) the Tennessee class size experiment, Project STAR (e.g., Finn and Achilles, 1990) and (2) the federal Reading First 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 Reading First Impact 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.

*Project STAR*

Project STAR (Student-Teacher Achievement Ratio) randomized approximately 5,900 entering kindergarten students at 79 elementary schools to either: a small class (13 – 17 students), a normal-size class (22 – 26 students) (Krueger, 1999 and Nye, Hedges and Konstantopoulos, 2000).[17] 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. We use 79 instruments—a zero/one indicator for assignment to a small class interacted with a zero/one indicator for each school. We use OLS, 2SLS with 79 instruments and the bias-corrected IV estimator 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*

[17] 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.

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 which 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 estimator, 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 $CV_\gamma$, $\tau_\delta$, and $Corr(\gamma, \delta)$ to determine whether, based on our simulations reported in Table 2 and Figures 2-3 above, we expect the bias-corrected estimator to outperform 2SLS. For both math and reading, $CV_\gamma \approx 0.25$ and $F > 1{,}000$.[18] Using the methods described in Raudenbush, Reardon, and Nomi (2012) and in in Appendix D, we estimate $\tau_\delta \approx 3.5$ for both math and reading, and $Corr(\gamma, \delta) = -0.24$ and $-0.36$ in math and reading, respectively. These values suggest that the bias-corrected estimator should perform extremely well. Based on Figure 2, when $CV_\gamma = 0.2$ and $Corr(\gamma, \delta) = 0.25$, the bias-corrected estimator is substantially less biased and has 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 bias-corrected IV estimates (reported at the bottom of Table 3) are larger (18 and 34 percent larger, respectively) than their 2SLS counterparts. They imply that reducing the size of a

[18] The large first-stage F-statistic reflects the facts that variation in class size is largely due to randomization and that the average sample per school is substantial.

kindergarten class by one student increases average student achievement by 1.32 scale-score points for math and 0.96 scale-score points for reading. Expressed as effect sizes these results are 0.028 and 0.031 standard deviations, respectively, per student of class size reduced. Note, however, that the standard errors of the bias-corrected estimates are 15-20% larger than the 2SLS and OLS 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 per block and numerous blocks, the three 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 we estimator we use in this case, the values of $CV_\gamma$, $F$, and $Corr(\gamma, \delta)$ and the simulations in Section V suggest that the bias-corrected estimates are to be preferred in this example.

Another potential way to assess the impact of compliance-effect covariance bias is to examine the estimates of $\alpha_1$. 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 ($\beta$) and the effect of randomization on class size ($\gamma$). This departure from linearity implies the presence of compliance-effect covariance bias.

To help visualize this relationship, Figure 4 presents a graph of reduced-form OLS estimates of $\beta$ and Empirical Bayes estimates of $\gamma$ for each school in the sample. Superimposed on this scatter-plot is the estimated quadratic relationship implied by the estimates of $\alpha_0$ and $\alpha_1$ in Table 3. The top graph is for reading and the bottom graph is for math. 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 non-linearity in the underlying relationship between class size and achievement.

*Reading First*

The Reading First 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.[19] 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 one and two. 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 Reading First is schools, 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 report in the introduction above, an IV analysis with a single instrument indicates that on average, student reading achievement increased by 0.37 scale score points ($\frac{4.29}{11.6}$) per additional minute of RF instruction. The right side of table 3 reports corresponding results obtained from OLS, 2SLS with multiple instruments and the bias-corrected IV estimator with multiple instruments. The OLS estimates indicate a very small mediator effect: an additional 0.037 or 0.122

[19] Treatment was not assigned randomly in most of the RF sites, but was rather assigned on the basis of an observed rating score. Our analysis here, like the impact analysis reported by Gamse et al (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.

scale score points per minute of RF instruction per daily reading block (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, respectively, estimates that are very close to the single-instrument estimate of 0.37 points per minute of RF instruction.[20]

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 estimator.[21] This suggests that the Reading First 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 6 schools) do 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. [22] 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

[20] Expressed as effect sizes, the OLS estimates imply an increase in reading achievement of 0.03 or 0.01 standard deviations per 10 minutes per day of RF instruction. The 2SLS estimates imply an effect size of 0.09 standard deviations for 18 or 36 blocks. We use the standard deviation of the test's national norming sample (44.75) to compute effect sizes.

[21] Table 3 indicates that the first stage *F*-statistic is 17.7 for 18 blocks and 8.2 for 36 blocks. Thus, reducing the number of instruments by aggregating classrooms across first and second grades increases the strength of the resulting instruments and thereby reduces finite sample bias.

[22] For the 2SLS and OLS estimators, it is possible to obtain estimated standard errors through conventional methods based on standard software packages. However, as demonstrated earlier in the paper, 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.

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 60–90 minute reading block). Neither 2SLS and nor the bias-corrected estimator are subject to this problem.

In summary, Project STAR illustrates a situation in which the bias-corrected estimator is likely to work quite well: its $F$-statistic is unusually large (over 1,000), its coefficient of variation for compliance equals about 0.26, and its number of observations per block (over 70) is large enough to support accurate bootstrapped standard errors. Reading First provides a much more limited application. Its $F$-statistic is 17.7 or 8.2, its coefficient of variation for compliance is 0.76 or 0.79 and its 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 multi-site 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 instrumental variables 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 and Raudenbush, forthcoming). This assumption is required regardless of whether the goal is to identify a complier average causal effect (a LATE, in Angrist, Imbens, and Rubin's 1996 terminology) or an average effect in a population (ATE).

Reardon and Raudenbush (forthcoming, Appendix C) 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 finite sample bias. 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 finite sample bias and compliance-effect covariance bias. This expression shows that, *ceteris parabis*, the magnitude of compliance-effect covariance bias increases asymptotically as the instruments grow stronger, while finite sample bias 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 an alternative instrumental variables estimator—the bias-corrected IV estimator. Our simulations show that this estimator performs very well over a wide range of conditions when the first stage $F$-statistic is greater than 10. In this situation, as long as $CV_\gamma$ is not too extreme and $Corr(\gamma_s, \delta_s)$ is not very close to zero, the bias-corrected estimator outperforms 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 estimator.

The bias-corrected estimator relies 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 estimator requires only that 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 estimator is therefore preferable to 2SLS in a wide range of situations for the analysis of mediator effects in multi-site trials.

**References**


Angrist, J. and Pischke, J. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review* 80(3): 313-336.

Borjas, G. J. (1987). "Self-Selection and the Earnings of Immigrants." *The American Economic Review* 77(4): 531-553.

Bound, J., A. Jaeger, and R. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90: 443-450.

Chamberlain, G. and Imbens, G. (2004). Random Effects Estimators with Many Instrumental Variables. *Econometrica*, Vol 72, No. 1, 295-306.

Duncan, G. J., Morris, P., and 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*, 48(5): 1263-1279.

Finn, J., and 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.* 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.

Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Hahn, Jinyong, and Jerry Hausman. (2002). "A New Specification Test for the Validity of Instrumental Variables." *Econometrica* 70:1: 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*(4), 974-987.

Johnson, N.L., & Kotz, S. (1994). *Continuous Univariate Distributions*, New York: Wiley.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica, 75*(1), 83-119.

Krueger, A. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 114(2), 497-532.

Nomi, T, and Raudenbush, S. (2012). The impact of math curricular reform on course-taking, classroom composition and achievement: A multi-site discontinuity design. Working Paper.

Nye, B., Hedges, L., and Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37(1), 123-151.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (*2nd ed*.).* Newbury Park, CA: Sage Publications.

Raudenbush, S., Reardon, S., and Nomi, T. (2012). Statistical Analysis for Multi-site Trials Using Instrumental Variables. *Journal or Research on Educational Effectiveness,* 5(3), 303-332.

Reardon, S. and Raudenbush, S. (forthcoming). Under What Assumptions do Site by Treatment Instruments Identify Average Causal Effects? *Sociological Methods and Research*.

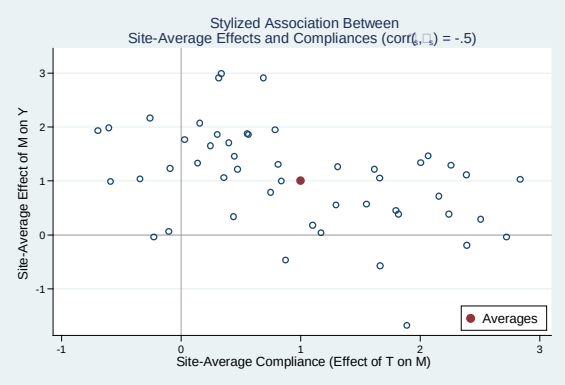
Roy, A.D. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* (New Series) 3 (1951): 135-146.

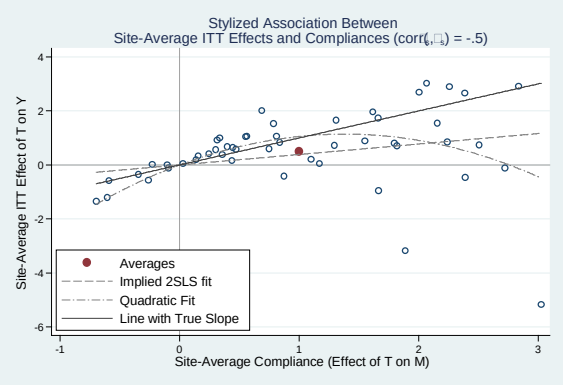
Staiger, D and Stock, J. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica* 65, no. 3 (1997), 557-586

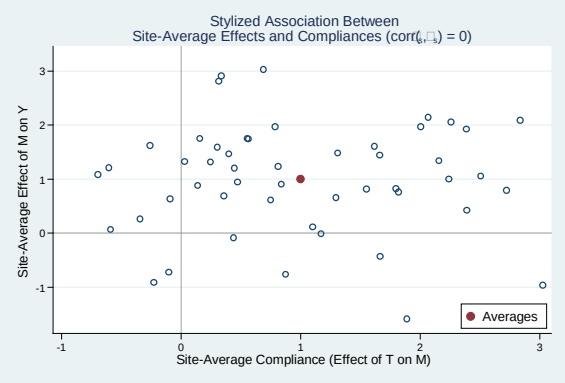
StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

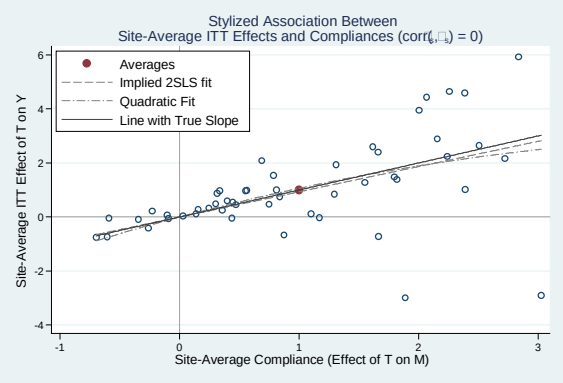
Stock, J. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression , *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (Donald W. K. Andrews and James H. Stock, eds.), Cambridge: Cambridge University Press, 80–108.

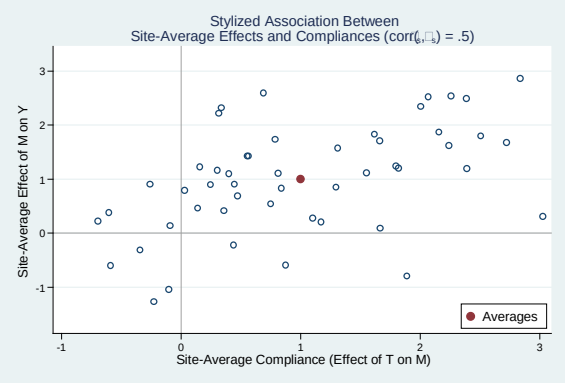
**Figure 1**


Stylized Association Between
Site-Average Effects and Compliances (corr(,) = -.5)
Site-Average Effect of M on Y
Site-Average Compliance (Effect of T on M)
Averages


Stylized Association Between
Site-Average ITT Effects and Compliances (corr(,) = -.5)
Site-Average ITT Effect of T on Y
Site-Average Compliance (Effect of T on M)
Averages
Implied 2SLS fit
Quadratic Fit
Line with True Slope


Stylized Association Between
Site-Average Effects and Compliances (corr(,) = 0)
Site-Average Effect of M on Y
Site-Average Compliance (Effect of T on M)
Averages


Stylized Association Between
Site-Average ITT Effects and Compliances (corr(,) = 0)
Site-Average ITT Effect of T on Y
Site-Average Compliance (Effect of T on M)
Averages
Implied 2SLS fit
Quadratic Fit
Line with True Slope


Stylized Association Between
Site-Average Effects and Compliances (corr(,) = .5)
Site-Average Effect of M on Y
Site-Average Compliance (Effect of T on M)
Averages

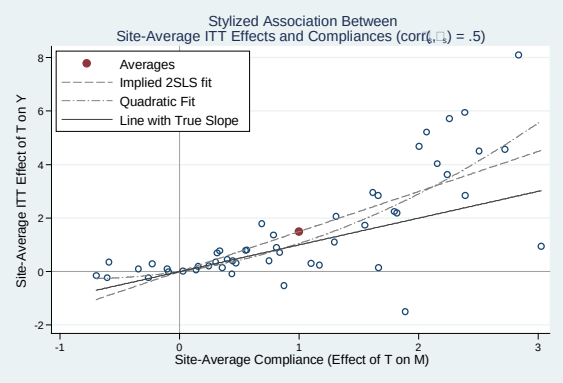

Stylized Association Between
Site-Average ITT Effects and Compliances (corr(,) = .5)
Site-Average ITT Effect of T on Y
Site-Average Compliance (Effect of T on M)
Averages
Implied 2SLS fit
Quadratic Fit
Line with True Slope

**Figure 2: Bias and RMSE of Three Estimators by F-statistic and $CV_\gamma$, when $Corr(\gamma_s, \delta_s)$= 0.25**

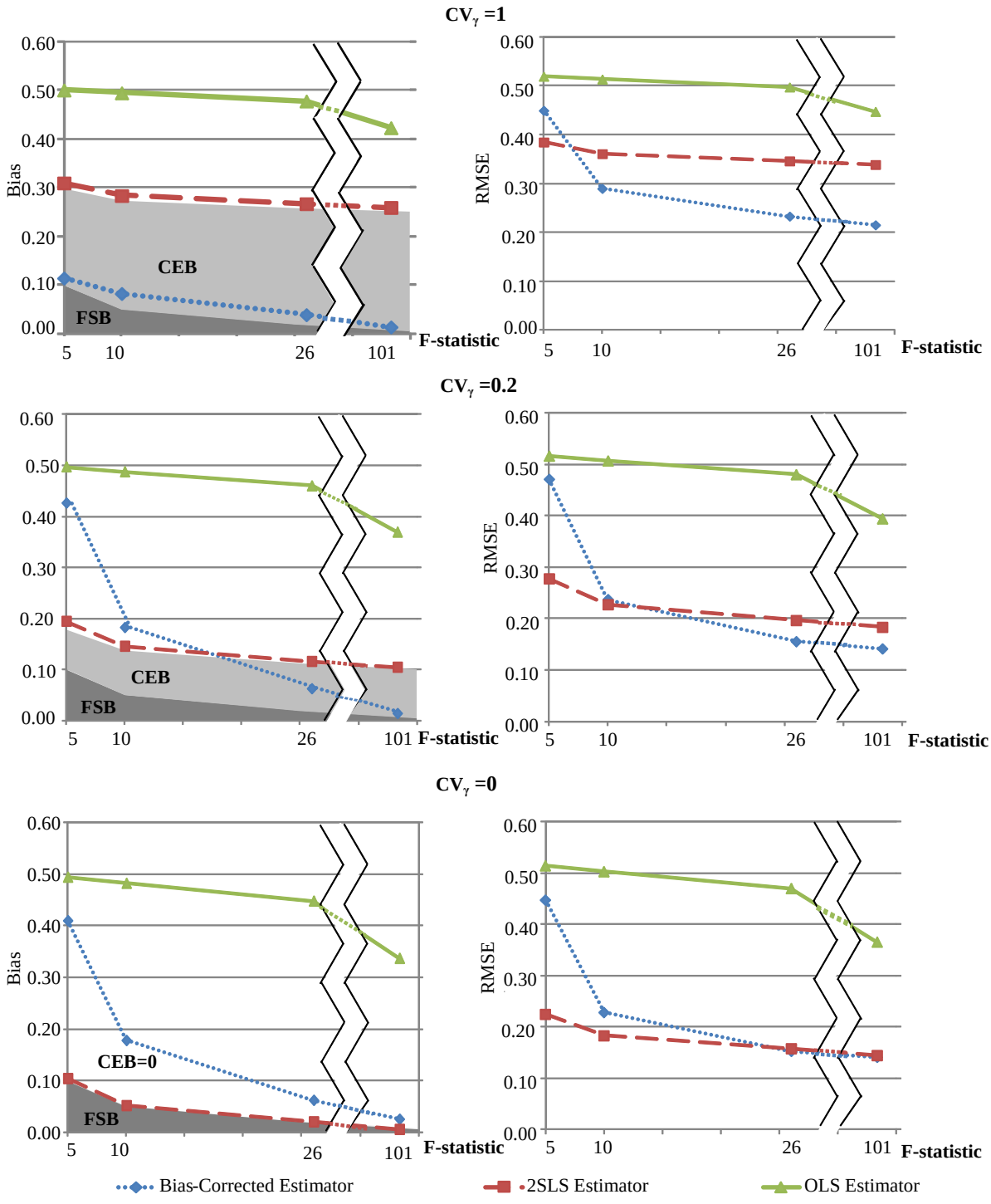


**Figure 3: Bias and RMSE of Three Estimators, by $\mathrm{Corr}(\gamma_s, \delta_s)$ and $CV_\gamma$, when F-statistic=26**

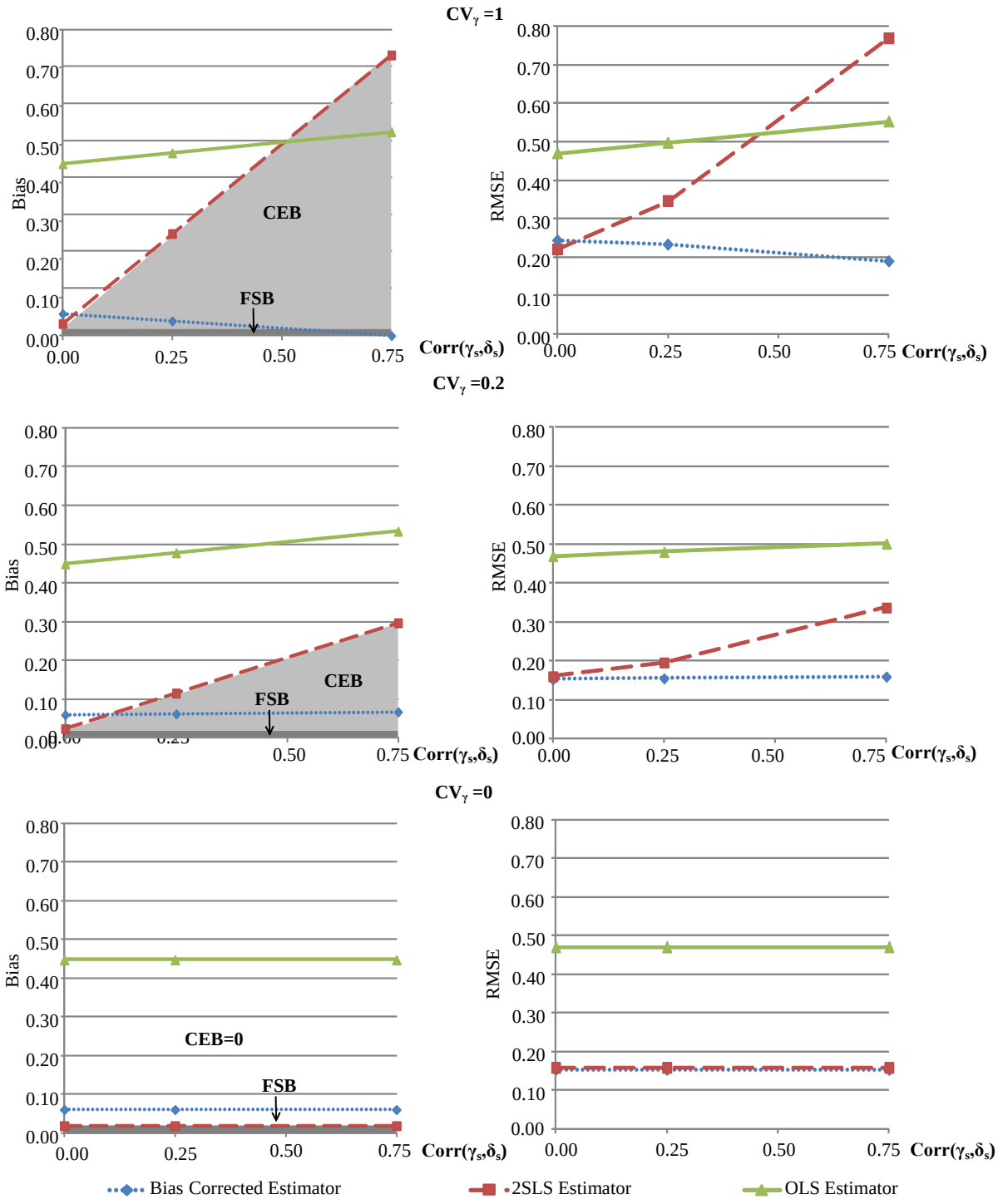
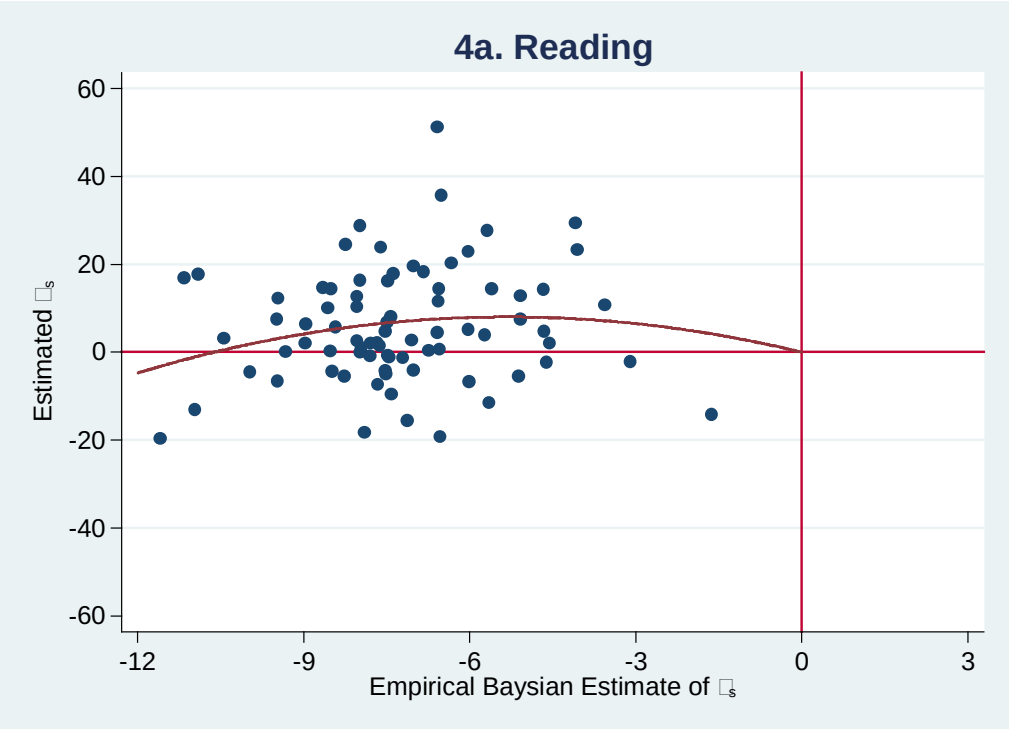


**Figure 4. Relationship between Reduced-form OLS Estimates of $\beta_s$ and Empirical Bayes Estimates of $\gamma_s$ for Each School in the Tennessee STAR Sample, for Kindergarten Reading and Math Test Scores**



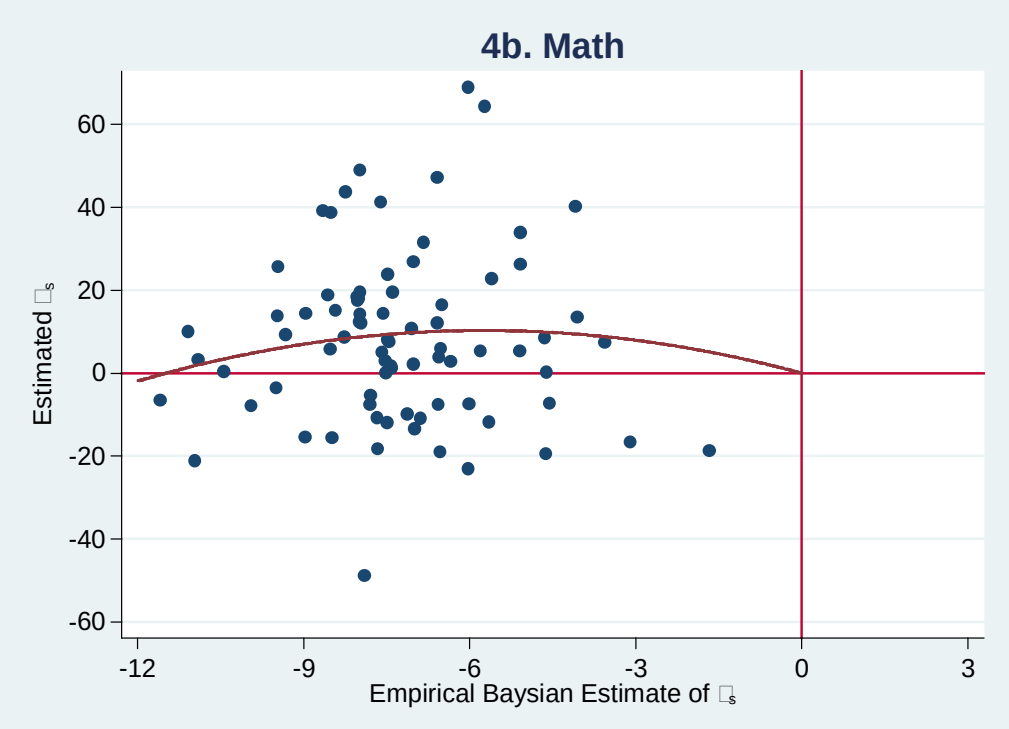



**Table 1. Estimated Bias and Root Mean Squared Error of Multiple-Site, Multiple-Instrument 2SLS Estimator**

| | Data Generating Parameters | | | | 2SLS Estimator | | | | | OLS Estimator | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | $CV_\gamma$ | $F$ | $Corr(\gamma_s, \delta_s)$ | sd (δ) | Predicted Bias | Estimated Bias | True $se(\hat{\delta})$ | Average $se(\hat{\delta})$ | RMSE | Predicted Bias | Estimated Bias | True $se(\hat{\delta})$ | Average $se(\hat{\delta})$ | RMSE |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) | (14) |
| **Panel A: $CV_\gamma$ varies** | | | | | | | | | | | | | | |
| 1 | 0 | 10 | 0.25 | 1 | 0.05 | 0.05 | 0.173 | 0.064 | 0.180 | 0.48 | 0.48 | 0.142 | 0.013 | 0.503 |
| 2 | 0.2 | 10 | 0.25 | 1 | 0.14 | 0.14 | 0.173 | 0.061 | 0.221 | 0.48 | 0.49 | 0.139 | 0.013 | 0.506 |
| 3 | 1 | 10 | 0.25 | 1 | 0.28 | 0.28 | 0.223 | 0.061 | 0.361 | 0.49 | 0.49 | 0.139 | 0.013 | 0.513 |
| 4 | 5 | 10 | 0.25 | 1 | 0.14 | 0.15 | 0.256 | 0.062 | 0.297 | 0.48 | 0.49 | 0.139 | 0.013 | 0.507 |
| 5 | ∞ | 10 | 0.25 | 1 | 0.05 | 0.07 | 0.259 | 0.064 | 0.270 | 0.48 | 0.48 | 0.140 | 0.013 | 0.504 |
| **Panel B: Expected *F*-statistic varies** | | | | | | | | | | | | | | |
| 6 | 1 | 2 | 0.25 | 1 | 0.38 | 0.39 | 0.243 | 0.135 | 0.457 | 0.50 | 0.50 | 0.139 | 0.013 | 0.523 |
| 7 | 1 | 5 | 0.25 | 1 | 0.30 | 0.31 | 0.229 | 0.086 | 0.385 | 0.50 | 0.50 | 0.139 | 0.013 | 0.519 |
| 8 | 1 | 10 | 0.25 | 1 | 0.28 | 0.28 | 0.223 | 0.061 | 0.361 | 0.49 | 0.49 | 0.139 | 0.013 | 0.513 |
| 9 | 1 | 26 | 0.25 | 1 | 0.26 | 0.27 | 0.220 | 0.039 | 0.346 | 0.47 | 0.48 | 0.139 | 0.013 | 0.497 |
| 10 | 1 | 101 | 0.25 | 1 | 0.25 | 0.26 | 0.218 | 0.021 | 0.339 | 0.42 | 0.42 | 0.146 | 0.013 | 0.447 |
| **Panel C: Corr($\gamma_s$,$\delta_s$) varies** | | | | | | | | | | | | | | |
| 11 | 1 | 10 | -0.75 | 1 | -0.63 | -0.60 | 0.240 | 0.078 | 0.649 | 0.45 | 0.45 | 0.145 | 0.013 | 0.469 |
| 12 | 1 | 10 | -0.25 | 1 | -0.18 | -0.16 | 0.225 | 0.067 | 0.275 | 0.47 | 0.47 | 0.139 | 0.013 | 0.491 |
| 13 | 1 | 10 | 0 | 1 | 0.05 | 0.06 | 0.223 | 0.063 | 0.232 | 0.48 | 0.48 | 0.138 | 0.013 | 0.502 |
| 14 | 1 | 10 | 0.25 | 1 | 0.28 | 0.28 | 0.223 | 0.061 | 0.361 | 0.49 | 0.49 | 0.139 | 0.013 | 0.513 |
| 15 | 1 | 10 | 0.75 | 1 | 0.73 | 0.72 | 0.234 | 0.061 | 0.757 | 0.51 | 0.52 | 0.142 | 0.013 | 0.536 |
| **Panel D: sd (δ) varies** | | | | | | | | | | | | | | |
| 16 | 1 | 10 | 0.25 | 0 | 0.05 | 0.05 | 0.045 | 0.044 | 0.068 | 0.48 | 0.48 | 0.009 | 0.010 | 0.479 |
| 17 | 1 | 10 | 0.25 | 0.2 | 0.10 | 0.10 | 0.061 | 0.044 | 0.114 | 0.48 | 0.48 | 0.028 | 0.012 | 0.483 |
| 18 | 1 | 10 | 0.25 | 1 | 0.28 | 0.28 | 0.223 | 0.061 | 0.361 | 0.49 | 0.49 | 0.139 | 0.013 | 0.513 |
| 19 | 1 | 10 | 0.25 | 5 | 1.18 | 1.22 | 1.101 | 0.234 | 1.644 | 0.53 | 0.56 | 0.696 | 0.013 | 0.892 |

Note: Details of simulation in Appendix B. In each row, the following parameters are used: Each simulation data sets 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 0.5. 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).

**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 | | | | 2SLS Estimator | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | CVγ | *F* | $Corr(\gamma_s,\delta_s)$ | Estimated Bias | True $se(\hat{\delta})$ | Average $se(\hat{\delta})$ | RMSE | Estimated Bias | True $se(\hat{\delta})$ | Average $se(\hat{\delta})$ | RMSE |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| **Panel A: $CV_\gamma$ varies** | | | | | | | | | | | |
| 1 | 0 | 10 | 0.25 | -0.18 | 0.14 | 0.16 | 0.23 | 0.05 | 0.17 | 0.06 | 0.18 |
| 2 | 0.2 | 10 | 0.25 | -0.18 | 0.15 | 0.17 | 0.24 | 0.14 | 0.17 | 0.06 | 0.22 |
| 3 | 1 | 10 | 0.25 | 0.08 | 0.28 | 0.27 | 0.29 | 0.28 | 0.22 | 0.06 | 0.36 |
| 4 | 5 | 10 | 0.25 | 0.19 | 0.25 | 0.22 | 0.31 | 0.15 | 0.26 | 0.06 | 0.30 |
| 5 | ∞ | 10 | 0.25 | 0.19 | 0.25 | 0.23 | 0.32 | 0.07 | 0.26 | 0.06 | 0.27 |
| **Panel B: Expected *F*-statistic varies** | | | | | | | | | | | |
| 6 | 1 | 2 | 0.25 | -0.35 | 2.66 | 0.77 | 2.69 | 0.39 | 0.24 | 0.13 | 0.46 |
| 7 | 1 | 5 | 0.25 | 0.11 | 0.43 | 0.36 | 0.45 | 0.31 | 0.23 | 0.09 | 0.38 |
| 8 | 1 | 10 | 0.25 | 0.08 | 0.28 | 0.27 | 0.29 | 0.28 | 0.22 | 0.06 | 0.36 |
| 9 | 1 | 26 | 0.25 | 0.04 | 0.23 | 0.22 | 0.23 | 0.27 | 0.22 | 0.04 | 0.35 |
| 10 | 1 | 101 | 0.25 | 0.01 | 0.21 | 0.21 | 0.22 | 0.26 | 0.22 | 0.02 | 0.34 |
| **Panel C: Corr($\gamma_s$,$\delta_s$) varies** | | | | | | | | | | | |
| 11 | 1 | 26 | 0.00 | 0.06 | 0.24 | 0.23 | 0.24 | 0.03 | 0.22 | 0.04 | 0.23 |
| 12 | 1 | 26 | 0.25 | 0.04 | 0.23 | 0.22 | 0.23 | 0.27 | 0.22 | 0.04 | 0.36 |
| 13 | 1 | 26 | 0.75 | 0.00 | 0.19 | 0.19 | 0.19 | 0.73 | 0.23 | 0.04 | 0.76 |

Note: Details of simulation in Appendix B. In each row, δ=1 and sd(δ)=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 2000 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) and (11) are computed as described in Table 1.

**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 s.e.(δ) | (0.340) | (0.230) | (n.a.) | (n.a.) |
| **2SLS Estimator** | | | | |
| δ | -1.114 ** | -0.714 ** | 0.397 | 0.387 |
| Bootstrapped s.e.(δ) | (0.350) | (0.230) | (n.a.) | (n.a.) |
| **Observable/Estimatble Parameters** | | | | |
| *F*-statistic | 1082.1 | 1071.5 | 17.7 | 8.2 |
| $\tau_\gamma$ | 3.45 | 3.47 | 63.25 | 68.30 |
| γ | -7.25 | -7.26 | 10.47 | 10.45 |
| $CV_\gamma$ | 0.26 | 0.26 | 0.76 | 0.79 |
| Estimated $\tau_\delta$ | 5.814 | 2.216 | 0.531 | 0.363 |
| Estimated Corr($\gamma_s$, $\delta_s$) | -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** | | | | |
| $\alpha_0$ | -3.583 * | -3.025 ** | 0.157 | 0.491 |
| s.e.($\alpha_0$) | (1.546) | (0.959) | (1.025) | (0.783) |
| $\alpha_1$ | -0.312 + | -0.285 * | 0.020 | -0.001 |
| s.e.($\alpha_1$) | (0.187) | (0.116) | (0.060) | (0.045) |
| **Bias-Corrected Estimator** | | | | |
| δ | -1.319 ** | -0.957 *** | 0.365 | 0.484 |
| Bootstrapped s.e.(δ) | (0.420) | (0.260) | (n.a.) | (n.a.) |
| N (sites/blocks) | 79 | 79 | 18 | 36 |
| N(observations) | 5871 | 5789 | 248 | 248 |

Note: + p<.10; * p<.05; ** p<.01; *** p<.001. Estimated compliance effect covariance bias computed from Equation (5a). Boostrapped standard errors computed as described in text.

**Appendix A**

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

$$M_i = \Lambda_s + \gamma_s T_i + e_i, \quad e_i \sim N(0, \sigma^2)$$

$$Y_i = \Theta_s + \delta_s M_i + u_i, \quad u_i \sim N(0, \omega^2)$$

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\omega \\ \rho\sigma\omega & \omega^2 \end{pmatrix} \right],$$

where $\rho$ is the within-site correlation of $e_i$ and $u_i$. Across sites, the covariance matrix of the $\gamma_s$'s and the $\delta_s$'s is

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

(1)

**A1: Derivation of the population *F*-statistic (Equation 3)**[23]

Suppose $W$ is distributed as a non-central chi-square with $df = \nu_1$ and non-centrality parameter $\lambda$; and $U$ is distributed as a central chi-square with $df = \nu_2$ independently of $W$, then $F = \frac{W/\nu_1}{U/\nu_2}$ will be distributed as $F(\nu_1, \nu_2, \lambda)$, a non-central $F$ with numerator degrees of freedom $\nu_1$, denominator degrees of freedom $\nu_2$ and non-centrality parameter $\lambda$. This variable has mean (Johnson & Kotz, 1994)

$$E[F] = \frac{\nu_2}{(\nu_2 - 2)} \cdot \left(1 + \frac{\lambda}{\nu_1}\right).$$

(A1.1)

When $\nu_2$ is large, $\frac{\nu_2}{(\nu_2 - 2)} \approx 1$, so we have

$$E[F] \approx 1 + \frac{\lambda}{\nu_1}.$$

(A1.2)

[23] We thank Steve Raudenbush for providing this derivation.

Now consider the data generating model given above. Define $z_s = \frac{\hat{\gamma}_s}{se(\hat{\gamma}_s)} = \frac{\bar{M}_s^{t=1} - \bar{M}_s^{t=0}}{\sigma/\sqrt{np(1-p)}}$, the ratio of the sample mean difference between experimental and control groups in site *s* and its standard error. Then $z_s$ is distributed as a non-central $Z$ with non-centrality parameter $\lambda_s = E[z_s] = \frac{\gamma_s}{\sigma/\sqrt{np(1-p)}}$. It follows that $W = \sum_{s=1}^{K} z_s^2$ is distributed as a non-central chi-square with degrees of freedom $\nu_1 = K$ and non-centrality parameter

$$\lambda = E\left[\sum_{s=1}^{K} \lambda_s^2\right] = E\left[\frac{\sum_{s=1}^{K} \lambda_s^2}{\sigma^2/np(1-p)}\right] = \frac{Knp(1-p)(\gamma^2 + \tau_\gamma)}{\sigma^2}.$$

(A1.3)

Now define

$$U = \sum_{s=1}^{K}\sum_{i=1}^{n}\left(M_{is} - \hat{\Lambda}_s - \hat{\gamma}_s T_{is}\right)^2/\sigma^2.$$

(A1.4)

$U$ is distributed as a central chi-square with $df = \nu_2 = K(n-2)$. Now note that $F = \frac{W/\nu_1}{U/\nu_2}$ is the $F$-statistic for the test of the null hypothesis that the instrument has no effect in every site, $H_0: \gamma_s = 0, \forall s$, or, alternately, $H_0: \sum_{s=1}^{K} \gamma_s^2 = 0$. So long as $\nu_2 = K(n-2)$ is large, Equation (A1.2) yields Equation (3):

$$E[F] \approx 1 + \frac{\lambda}{\nu_1} = 1 + \frac{np(1-p)\left(\gamma^2 + \tau_\gamma\right)}{\sigma^2}$$

(3)

**A2: Derivation of OLS bias (Equation 4a)**

Let $X_i^+ = X_i - \bar{X}_s$ denote the within-site centered value of a variable $X$. Then centering both sides of Equation (2b) and substituting in the centered version of (2a) yields

$$
\begin{aligned}
Y_i^+ &= \delta_s M_i^+ + u_i^+ \\
&= \delta M_i^+ + [(\delta_s - \delta) M_i^+ + u_i^+] \\
&= \delta M_i^+ + [(\delta_s - \delta)(\gamma_s T_i^+ + e_i^+) + u_i^+] \\
&= \delta M_i^+ + [(\delta_s - \delta)(\gamma T_i^+ + (\gamma_s - \gamma) T_i^+ + e_i^+) + u_i^+].
\end{aligned}
\tag{A2.1}
$$

Estimating $\delta$ via OLS yields

$$
\begin{aligned}
E\left[\hat{\delta}^{OLS}\right] &= \frac{E[Cov(Y_i^+, M_i^+)]}{Var\left(M_i^+\right)} \\
&= \frac{E[Cov(\delta M_i^+ + [(\delta_s - \delta)(\gamma T_i^+ + (\gamma_s - \gamma) T_i^+ + e_i^+) + u_i^+], M_i^+)]}{Var\left(M_i^+\right)} \\
&= \delta + \frac{E\left[Cov\left([(\delta_s - \delta)(\gamma T_i^+ + (\gamma_s - \gamma) T_i^+ + e_i^+) + u_i^+], (\gamma T_i^+ + (\gamma_s - \gamma) T_i^+ + e_i^+)\right)\right]}{Var\left(\gamma T_i^+ + (\gamma_s - \gamma) T_i^+ + e_i^+\right)} \\
&= \delta + \frac{E\left[2\gamma \tau_{\gamma\delta} Var(T_i^+) + Cov(u_i^+, e_i^+)\right]}{\gamma^2 Var\left(T_i^+\right) + \tau_\gamma Var\left(T_i^+\right) + Var\left(e_i^+\right)} \\
&= \delta + \frac{2p(1-p)\gamma\tau_{\gamma\delta} + \rho\omega\sigma}{p(1-p)\left(\gamma^2 + \tau_\gamma\right) + \sigma^2} \\
&= \delta + \frac{2\frac{np(1-p)}{\sigma^2}\gamma\tau_{\gamma\delta} + \frac{n\rho\omega}{\sigma}}{\frac{np(1-p)}{\sigma^2}\left(\gamma^2 + \tau_\gamma\right) + n} \\
&= \delta + \frac{2\frac{np(1-p)}{\sigma^2}\gamma\tau_{\gamma\delta} + \frac{n\rho\omega}{\sigma}}{F + n - 1} \\
&= \delta + \rho\frac{\omega}{\sigma}\left(\frac{n}{F + n - 1}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\frac{F - 1}{F + n - 1}\right)
\end{aligned}
\tag{A2.2}
$$

**A3: Derivation of 2SLS bias (Equation 5a)**

Combining Equations (2a) and (2b) yields the reduced form Equation

$$
\begin{aligned}
Y_i &= \Theta_s + \delta_s \Lambda_s + \delta_s \gamma_s T_i + \delta_s e_i + u_i \\
&= \mathrm{A}_s + \beta_s T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \delta_s^2 \sigma^2 + \omega^2 + 2\delta_s \rho \sigma \omega)
\end{aligned}
$$

(A3.1)

We begin by fitting Equation (2a) via OLS. This yields estimates of the average compliance in each site $s$:

$$
\hat{\gamma}_s = \gamma_s + \nu_s, \quad \nu_s \sim N\left(0, \frac{\sigma^2}{np(1-p)}\right).
$$

(A3.2)

We also can estimate, within each site, the average ITT effect $\beta_s$. Here we have:

$$
\hat{\beta}_s = \beta_s + \eta_s, \quad \eta_s \sim N\left(0, \frac{(\delta_s^2 \sigma^2 + \omega^2 + 2\delta_s \rho \sigma \omega)}{np(1-p)}\right).
$$

(A3.3)

In finite samples:

$$
\begin{aligned}
\hat{\gamma}_s &= \gamma_s + g_s \\
\hat{\beta}_s &= \beta_s + b_s,
\end{aligned}
$$

(A3.4)

where $g_s = (\bar{e}_s^1 - \bar{e}_s^0)$ and $b_s = \delta_s(\bar{e}_s^1 - \bar{e}_s^0) + (\bar{u}_s^1 - \bar{u}_s^0)$, and where $\bar{e}_s^t$ and $\bar{u}_s^t$ are the average values of the error terms $e_i$ and $u_i$ among those with $T = t$ in the site $s$ sample. Now,

$$
\begin{aligned}
Cov(g_s, b_s) &= Cov\big(\bar{e}_s^1 - \bar{e}_s^0, \delta_s(\bar{e}_s^1 - \bar{e}_s^0)\big) + Cov(\bar{e}_s^1 - \bar{e}_s^0, \bar{u}_s^1 - \bar{u}_s^0) \\
&= \delta Var(\bar{e}_s^1 - \bar{e}_s^0) + Cov(\bar{e}_s^1 - \bar{e}_s^0, \bar{u}_s^1 - \bar{u}_s^0) \\
&= \delta\big(Var(\bar{e}_s^1) + Var(\bar{e}_s^0)\big) + Cov(\bar{e}_s^1, \bar{u}_s^1) + Cov(\bar{e}_s^0, \bar{u}_s^0) \\
&= \delta\left(\frac{\sigma^2}{pn} + \frac{\sigma^2}{(1-p)n}\right) + \left(\frac{\rho\sigma\omega}{pn} + \frac{\rho\sigma\omega}{(1-p)n}\right) \\
&= \frac{\delta\sigma^2 + \rho\sigma\omega}{np(1-p)}.
\end{aligned}
$$

(A3.5)

Under the assumption of no within-site compliance-effect covariance, $\beta_s = \gamma_s \delta_s$. Thus

$$Cov(\hat{\gamma}_s, \hat{\beta}_s) = Cov(\gamma_s, \beta_s) + Cov(g_s, b_s)$$

$$= Cov(\gamma_s, \gamma_s \delta_s) + \frac{\delta \sigma^2 + \rho \sigma \omega}{np(1-p)}$$

$$= \tau_\gamma \delta + \gamma Cov(\gamma_s, \delta_s) + \frac{\delta \sigma^2 + \rho \sigma \omega}{np(1-p)}.$$

(A3.6)

Note that 2SLS with site-by-treatment interactions is equivalent to fitting the regression model

$$\hat{\beta}_s = \delta \hat{\gamma}_s + v_s$$

(A3.7)

via WLS, weighting each point by $W_s = n_s p_s (1 - p_s)$. This yields

$$\hat{\delta}^{(2SLS)} = \frac{\sum_{s=1}^{K} n_s p_s (1-p_s) \hat{\gamma}_s \hat{\beta}_s}{\sum_{s=1}^{K} n_s p_s (1-p_s) \hat{\gamma}_s^2}$$

(A3.8)

Under the assumption that that $n_s = n$ and $p_s = p$ for all $s$, we have

$$\hat{\delta}^{(2SLS)} = \frac{\sum_{s=1}^{J} \hat{\gamma}_s \hat{\beta}_s}{\sum_{s=1}^{J} \hat{\gamma}_s^2}$$

(A3.9)

Now the expected value of the 2SLS estimator will be approximately equal to the ratio of the expected values of the numerator and denominator:

$$E\left[\hat{\delta}^{(2SLS)}\right] \approx \frac{E[\sum_{s=1}^{J} \hat{\gamma}_s \hat{\beta}_s]}{K\left(\gamma^2 + \tau_\gamma + \frac{\sigma^2}{np(1-p)}\right)}$$

$$= \frac{\gamma\beta + E\left[Cov(\hat{\gamma}_s, \hat{\beta}_s)\right]}{\left(\gamma^2 + \tau_\gamma + \frac{\sigma^2}{np(1-p)}\right)}$$

$$= \frac{\gamma(\gamma\delta + \tau_{\gamma\delta}) + \tau_\gamma\delta + \gamma\tau_{\gamma\delta} + \frac{\delta\sigma^2 + \rho\sigma\omega}{np(1-p)}}{\left(\gamma^2 + \tau_\gamma + \frac{\sigma^2}{np(1-p)}\right)}$$

$$= \delta + \frac{2\gamma\tau_{\gamma\delta} + \frac{\rho\sigma\omega}{np(1-p)}}{\left(\gamma^2 + \tau_\gamma + \frac{\sigma^2}{np(1-p)}\right)}$$

$$= \delta + \frac{2\gamma\tau_{\gamma\delta}}{\left(\gamma^2 + \tau_\gamma + \frac{\sigma^2}{np(1-p)}\right)} + \frac{\rho\sigma\omega}{np(1-p)(\gamma^2 + \tau_\gamma) + \sigma^2}$$

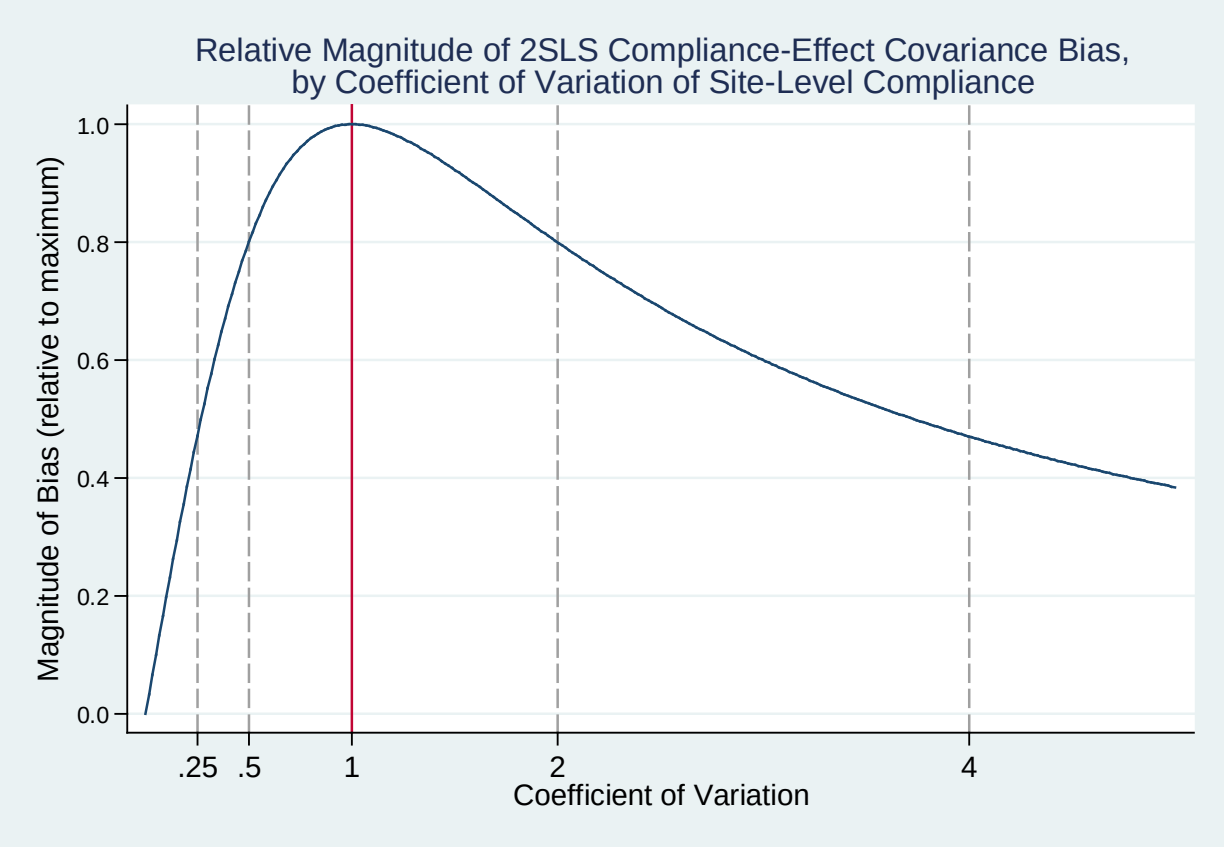
$$= \delta + \rho\frac{\omega}{\sigma}\left(\frac{1}{F}\right) + \frac{2\gamma\tau_{\gamma\delta}}{\gamma^2 + \tau_\gamma}\left(\frac{F-1}{F}\right)$$

(A3.10)

**A4: Proof that CEC bias is maximized when $CV_\gamma = 1$**

Equation (6) shows that compliance-effect covariance bias depends linearly on $\frac{CV_\gamma}{CV_\gamma^2+1}$. Let $f(x) = \frac{x}{x^2+1}$. Then note that $f\left(\frac{1}{x}\right) = \frac{\frac{1}{x}}{\frac{1}{x^2}+1} = \frac{x}{x^2+1} = f(x)$. We consider only the case where $x \geq 0$, because the sign of the $CV$ is arbitrary. A plot of $f(x)$ is shown below, indicating that $f(x)$ is maximized when $x = 1$. Note that for values of $x$ between 0.5 and 2, the bias is at least 80% of its maximum; for values less than 0.25 or greater than 4, the relative bias is less than half its maximum possible.

**Figure A4.1**




**A5: Derivation of Eq 13:**

Equation (12) indicates that $\beta_s$ can be written as a quadratic function of $\gamma_s$ plus a heteroskedastic error term:

$$\beta_s = \alpha_0\gamma_s + \alpha_1\gamma_s^2 + \gamma_s v_s \tag{A5.1}$$

Adding the sampling error in $\hat{\beta}_s$ to both sides of the equation yields

$$\hat{\beta}_s = \alpha_0\gamma_s + \alpha_1\gamma_s^2 + \gamma_s v_s + b_s. \tag{A5.2}$$

Taking the expectation, given the estimated $\hat{\gamma}_s$'s, yields

$$E[\hat{\beta}_s|\hat{\gamma}_s] = E[\alpha_0\gamma_s|\hat{\gamma}_s] + E[\alpha_1\gamma_s^2|\hat{\gamma}_s] + E[\gamma_s v_s|\hat{\gamma}_s] + E[b_s|\hat{\gamma}_s].$$

(A5.3)

Now define $\gamma_s^* = E[\gamma_s|\hat{\gamma}_s] = \lambda\hat{\gamma}_s + (1-\lambda)\gamma$, where $\lambda = \tau_\gamma/(\tau_\gamma + \tau_g)$ is the reliability of the $\hat{\gamma}_s$'s. In addition, define $\gamma_s^{2^*} = E[\gamma_s^2|\hat{\gamma}_s] = \gamma_s^{*2} + \tau_\gamma(1-\lambda)$. Then, noting that $v_s \perp \gamma_s$, we have

$$E\left[\hat{\beta}_s\middle|\hat{\gamma}_s\right] = \alpha_0\gamma_s^* + \alpha_1\gamma_s^{2^*} + E[b_s|\hat{\gamma}_s].$$

(A5.4)

Now, note that

$$\begin{aligned}
E[b_s|\hat{\gamma}_s] &= E[b_s|\hat{\gamma}_s = \gamma] + \frac{Cov(b_s,\hat{\gamma}_s)}{Var(\hat{\gamma}_s)}(\hat{\gamma}_s - \gamma) \\
&= E[b_s] + \frac{Cov(b_s,g_s)}{\tau_\gamma + \tau_g}(\hat{\gamma}_s - \gamma) \\
&= Cov(b_s,g_s)\frac{\lambda(\hat{\gamma}_s - \gamma)}{\tau_\gamma} \\
&= Cov(b_s,g_s)\frac{\lambda\hat{\gamma}_s + (1-\lambda)\gamma - \gamma}{\tau_\gamma} \\
&= \frac{Cov(b_s,g_s)}{\tau_\gamma}\gamma_s^* - \frac{Cov(b_s,g_s)}{\tau_\gamma}\gamma
\end{aligned}$$

(A5.5)

Substituting this into (A5.4) and rearranging, we have

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

(A5.6)

This indicates that if we fit the model

$$\hat{\beta}_s = c + a_0\gamma_s^* + a_1\gamma_s^{2^*} + \eta_s,$$

(A5.7)

we will obtain

$$E[\hat{c}] = -\frac{\gamma Cov(b_s,g_s)}{\tau_\gamma}$$

$$E[\hat{a}_0] = \alpha_0 + \frac{Cov(b_s, g_s)}{\tau_\gamma}$$

$$E[\hat{a}_1] = \alpha_1$$

(A5.8)

Recall that, under our assumptions, $\delta = \alpha_0 + \alpha_1\gamma$. This suggests the following estimator for $\delta$:

$$\hat{\delta} = \hat{a}_0 + \frac{\hat{c}}{\hat{\gamma}} + \hat{a}_1\hat{\gamma}.$$

(A5.9)

If $\hat{\gamma}$ is reasonably precisely estimated, then

$$E[\hat{\delta}] = E[\hat{a}_0] + E\left[\frac{\hat{c}}{\hat{\gamma}}\right] + E[\hat{a}_1\hat{\gamma}]$$

$$\approx E[\hat{a}_0] + \frac{E[\hat{c}]}{E[\hat{\gamma}]} + E[\hat{a}_1]E[\hat{\gamma}]$$

$$\approx \alpha_0 + \frac{Cov(b_s, g_s)}{\tau_\gamma} - \frac{\gamma Cov(b_s, g_s)}{\tau_\gamma \gamma} + \alpha_1\gamma$$

$$\approx \alpha_0 + \alpha_1\gamma.$$

(A5.10)

However, if $\frac{Cov(b_s, g_s)}{\tau_\gamma}$ is small, then fitting A5.7 will yield

$$E[\hat{c}] \approx 0$$

$$E[\hat{a}_0] \approx \alpha_0$$

$$E[\hat{a}_1] = \alpha_1,$$

(A5.11)

which suggests that we can fit instead the model

$$\hat{\beta}_s = a_0\gamma_s^* + a_1{\gamma_s^2}^* + \eta_s,$$

(A5.12)

and instead estimate $\delta$ as:

$$\hat{\delta} = \hat{a}_0 + \hat{a}_1\hat{\gamma}.$$

(A5.13)

Although the latter model will yield biased estimates, it will be more efficient, and this efficiency gain may outweigh the bias (that is, this estimator in A5.13 may have smaller root mean squared error than that in A5.0).

Note that Equation A3.5 provides an expression for $Cov(b_s, g_s)$:

$$Cov(b_s, g_s) = \frac{\delta\sigma^2 + \rho\sigma\omega}{np(1-p)}$$

$$= \frac{\gamma^2 + \tau_\gamma}{F-1}\left(\delta + \rho\frac{\omega}{\sigma}\right).$$

(A5.14)

Thus, the expected value of the intercept in the regression model will be

$$E[\hat{c}] = \frac{-\gamma\left(\gamma^2 + \tau_\gamma\right)}{(F-1)\tau_\gamma}\left(\delta + \rho\frac{\omega}{\sigma}\right)$$

$$= \frac{-\gamma(1 + CV_\gamma^2)}{(F-1)CV_\gamma^2}\left(\delta + \rho\frac{\omega}{\sigma}\right).$$

(A5.15)

Note that the intercept will be large, in general, when $\gamma$ is large but $\tau_\gamma$ is small (i.e., when $CV_\gamma$ is small). However, in these cases, the sampling variance of both $\hat{c}$ and $\hat{\alpha}_0$ will be very large, as the regression model in A5.8 will have little variance in the $\gamma_s^*$'s (other than sampling variance, which will be non-informative) and estimation of $\hat{c}$ and $\hat{\alpha}_0$ will rely on substantial extrapolation. In contrast, when $\gamma$ is small and $\tau_\gamma$ is large (i.e., when $CV_\gamma$ is large), the intercept will be close to zero, in which case, fitting model A5.12 may be sufficient to provide an approximately unbiased estimate of $\delta$.

Because the estimator in A5.9 will be very imprecise in the cases when it is most needed (when $CV_\gamma$ is small and $F$ is small), we choose to use the estimator in A5.13 instead, as it has much less sampling variance than the former. In simulations not shown we confirmed that the A5.9

estimator has a larger root mean squared error (often extremely large) than that in A5.13. Based on this, we report results in the paper based on the no-intercept estimator.

**Appendix B. 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

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

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 ($\Lambda_s$ and $\Theta_s$ in our notation) with every other parameter to be zero. The means of $\Lambda_s$ and $\Theta_s$ 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 , $\tau_\gamma^2$, $\tau_\delta^2$, and $\tau_{\gamma\delta}$, we can set $CV_\gamma$, $F$, $Corr(\gamma_s, \delta_s)$, and $\sqrt{\tau_\delta^2}$, to the values used in Tables 1 and 2. Specifically, we set

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

These values ensure that the simulations correspond to the scenarios described in Tables 1 and 2.[24]

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

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

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:

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

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

**Appendix C: Additional Comparisons Among OLS, 2SLS, and the Bias-Corrected Estimator**

Figures 2 and 3 present simulation results for the OLS, 2SLS, and bias-corrected estimators as $CV_\gamma$ deviates from 1 towards 0. Figures C1 and C2 show the same results for the three estimators of interest as $CV_\gamma$ deviates from 1 towards infinity. Patterns observed in these cases closely mirror those in Figures 2 and 3.

[24] The 0.02 term in Equation (B.2) comes from the fact that we set $\sigma^2 = 1$ below.

**Figure C-1: Bias and RMSE of Three Estimators by F-statistic and $CV_\gamma$, when $Corr(\gamma_s, \delta_s)$= 0.25**

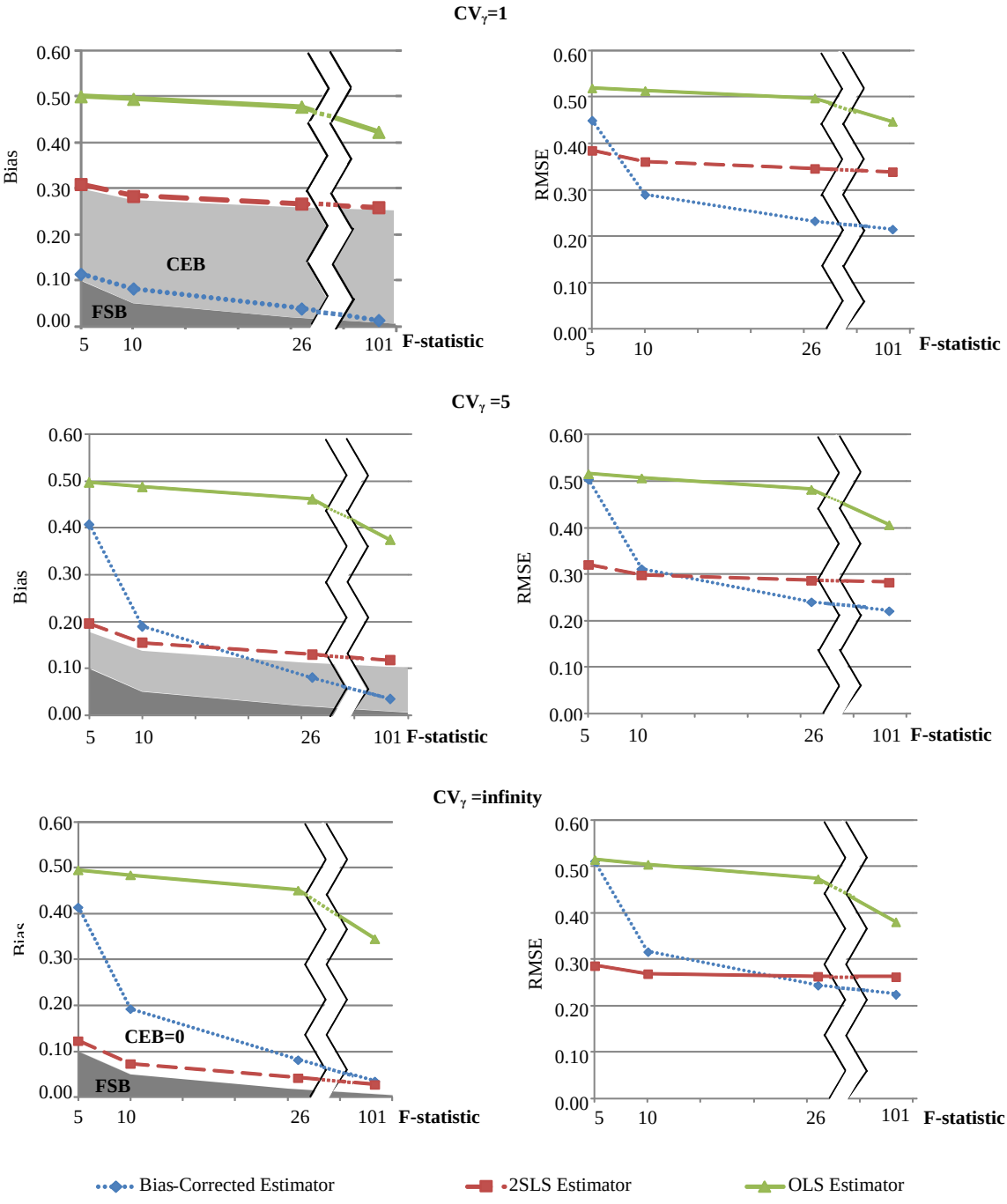
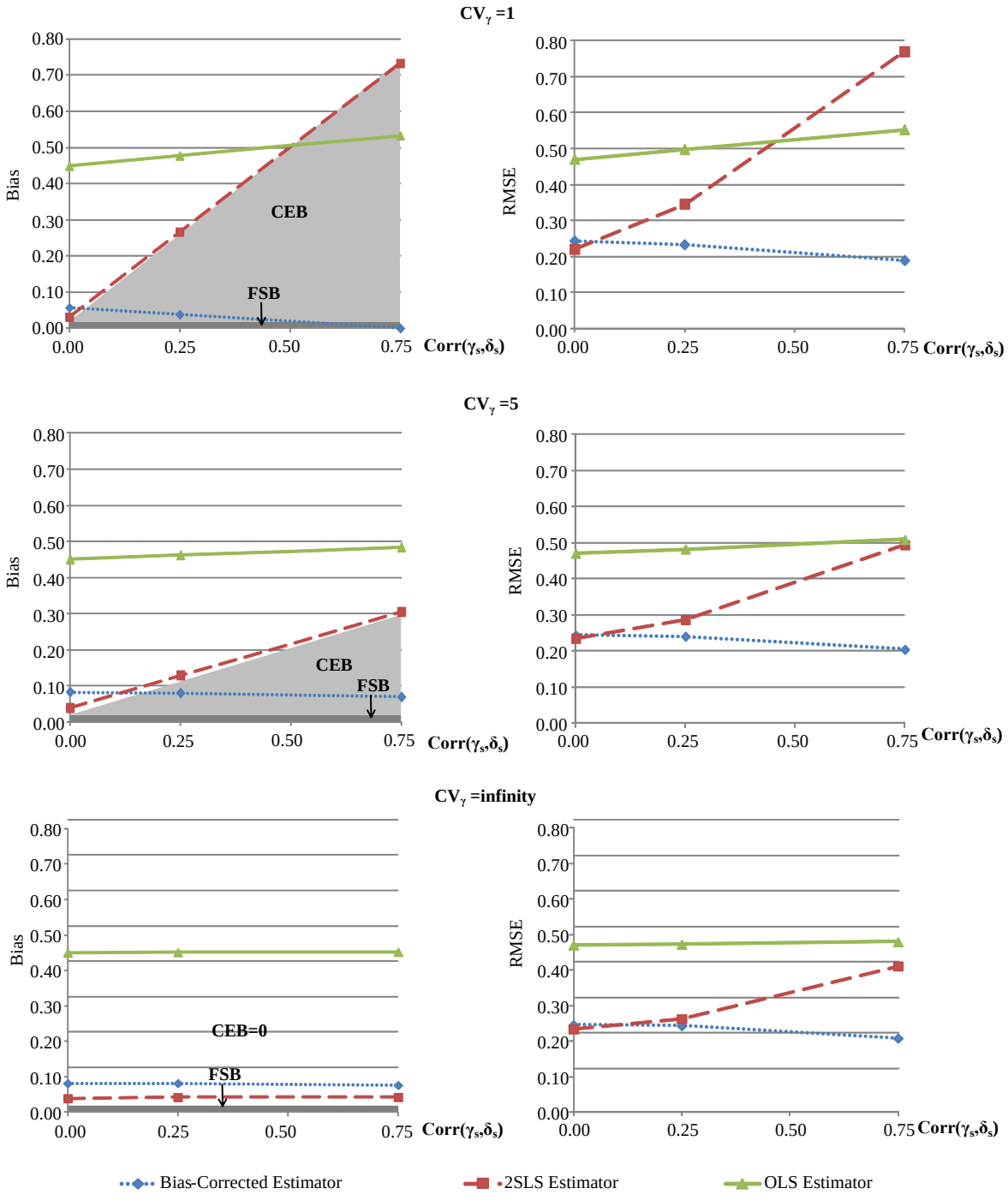


**Figure C-2: Bias and RMSE of Three Estimators, by $\mathrm{Corr}(\gamma_s, \delta_s)$ and $CV_\gamma$, when F-statistic=26**



**Appendix D: Estimating the correlation between $\gamma_s$ and $\delta_s$**

Equation (7) implies that

$$\frac{Cov(\gamma_s, \delta_s)}{\tau_\gamma} = \alpha_1$$

$$Cov(\gamma_s, \delta_s) = \alpha_1 \tau_\gamma$$

$$Corr(\gamma_s, \delta_s) = \alpha_1 \sqrt{\frac{\tau_\gamma}{\tau_\delta}}.$$

This implies that we can estimate $Corr(\gamma_s, \delta_s)$ if we can estimate $\alpha_1$, $\tau_\gamma$, and $\tau_\delta$ reasonably precisely. We obtain $\hat{\alpha}_1$ from fitting Equation (13), and we obtain $\tau_\gamma$from the random-coefficients first-stage model (Equation 10). Estimating $\tau_\delta$ is not as straightforward. We estimate $\tau_\delta$using the methods described in Raudenbush, Reardon, and Nomi (2012). The resulting estimates of $Corr(\gamma_s, \delta_s)$ are shown in Table 3.