Journal of Research on Educational Effectiveness

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uree20

Statistical Analysis for Multisite Trials Using Instrumental Variables With Random Coefficients

Stephen W. Raudenbush [a] , Sean F. Reardon [b] & Takako Nomi [c]

[a] The University of Chicago , Chicago , Illinois , USA

[b] Stanford University , Stanford , California , USA

[c] Consortium on Chicago School Research at The University of Chicago , Chicago , Illinois , USA

Published online: 12 Jul 2012.

To cite this article: Stephen W. Raudenbush , Sean F. Reardon & Takako Nomi (2012) Statistical Analysis for Multisite Trials Using Instrumental Variables With Random Coefficients, Journal of Research on Educational Effectiveness, 5:3, 303-332, DOI: 10.1080/19345747.2012.689610

To link to this article: http://dx.doi.org/10.1080/19345747.2012.689610

Journal of Research on Educational Effectiveness, 5: 303–332, 2012

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2012.689610

Routledge
Taylor & Francis Group

Statistical Analysis for Multisite Trials Using Instrumental Variables With Random Coefficients

Stephen W. Raudenbush
The University of Chicago, Chicago, Illinois, USA

Sean F. Reardon
Stanford University, Stanford, California, USA

Takako Nomi
Consortium on Chicago School Research at The University of Chicago, Chicago, Illinois, USA

Abstract: Multisite trials can clarify the average impact of a new program and the heterogeneity of impacts across sites. Unfortunately, in many applications, compliance with treatment assignment is imperfect. For these applications, we propose an instrumental variable (IV) model with person-specific and site-specific random coefficients. Site-specific IV coefficients can be interpreted as site-average effects of program participation or as site-average effects of participation for the compliers. The validity of these interpretations depends on the analyst's assumptions. Within the framework of a two-level hierarchical linear model, we propose three ways to estimate the mean and variance of these site-specific effects: (a) estimate the impact of program participation and its standard error in each site, then combine these site-specific statistics to estimate the mean and variance of the true site effects; (b) estimate the mean and variance of the effect of treatment assignment on the outcome and the mean and variance of the effect of treatment assignment on program participation; then combine these results to obtain estimates of the mean and variance of the effect of program participation; and (c) use Site by Treatment interactions as multiple instruments. If we assume the IV coefficients to be homogenous across sites, the three approaches are equivalent to variants of familiar two-stage least squares estimates with site fixed effects. Estimates based on our model are valid under a weaker assumption: that site-average levels of compliance are independent of site-average effects of program participation. To illustrate our approach, we evaluate a district-wide policy intended to increase math instructional time in math for low-achieving students. Finally, we discuss how Method (c) can be extended to incorporate multiple mediators.

Keywords: Instrumental variables, Multi-site trials, causal inference, multilevel analysis, high school curricular reform

In many large-scale program evaluations, the aim is to assess the impact of a novel program implemented within each of many program sites. For example, the recent national Head Start experiment (Administration for Children, Youth, and Families, 2010) sampled 380 program sites from a national list of all program sites; within each site, children were randomly assigned to attend Head Start. In the Tennessee class size experiment, teachers were assigned to a large or small class within each of 79 participating schools (Finn & Achilles, 1990). In fact, the vast majority of the 130 randomized field trials funded by the

Address correspondence to Stephen W. Raudenbush, Department of Sociology, 1126 East 59th Street, Chicago, IL 60637, USA. E-mail: sraudenb@uchicago.edu

Institute of Education Sciences over the past decade have been multisite trials (Spybrook & Raudenbush, 2009). In some of these studies, schools were assigned at random to treatments within each of several districts; in others, classrooms were assigned at random within each of many schools.

Multisite trials enable us to study the generalizability of a treatment effect across an array of varied contexts. Each site provides an independent trial of an intervention, enabling the researcher, in principle, to quantify the extent to which program impacts vary across settings and to model this variation. Bloom et al. (2011) noted that researchers have rarely exploited this potential and discussed methods for studying the variation in impact of treatment assignment in multisite randomized trials. The effect of treatment assignment, often called the "intent-to-treat" effect, is the difference between the average outcome of those participants assigned to a novel program and the average outcome those participants would have displayed if assigned to a counterfactual program, the control condition. The effect of treatment assignment is potentially important for policy but is not equivalent to the impact of actually participating in the program unless all participants "comply" with their treatment assignment, that is, everyone who is assigned to the novel program participates in that program and no one assigned to the control condition participates in the program (cf. Angrist, Imbens, & Rubin, 1996; Bloom, 1984).

In this article, we focus on estimating the effect of program participation when compliance is partial and may vary across sites in a multisite trial. In the National Head Start trial, many students assigned to Head Start did not in fact attend Head Start, and some of those assigned to the control condition showed up in a Head Start program. Thus, the average effect of being assigned to Head Start was by no means equivalent to the average effect of actually participating in Head Start. In the Tennessee class size study, students assigned to "small classes" found themselves in classes that varied quite considerably in size; similarly, not all students assigned to large classes actually experienced a large class. Although an important question is whether assignment to a small class affects outcomes, it is at least as interesting to know the impact on student learning of the realized class size (Krueger, 1999; Krueger & Whitmore, 2001, Nye, Konstantopoulos, & Hedges, 2004; Shin & Raudenbush, 2011).

The method of instrumental variables (IVs) is now almost universally recommended to study the impact of program participation in randomized experiments when compliance with randomization is imperfect: Treatment assignment is an "instrument" used to identify the impact of actually experiencing the program (cf. Angrist et al., 1996; Heckman & Robb, 1985). This can be regarded as a special case of a mediating model in which treatment assignment, to be designated as T in the current article, predicts program participation, designated here as M; there is an outcome of interest, designed as Y.

Researchers using the IV method have often implicitly assumed that everyone gains the same amount from participating in the program—or that persons within observable subgroups gain the same amount. However, economists often regard this assumption as naïve and instead assume that unobservable differences among participants are correlated with gains from program participation (Angrist et al., 1996; Heckman & Robb, 1985; Heckman & Vytlacil, 1998). Statisticians working within the counterfactual account of causality also assume treatment effects to be heterogeneous (Holland, 1986; Rubin, 1978) and have extended this idea to IV analysis (Angrist et al., 1996). In the case of multisite trials, heterogeneity in the impact of program participation has two sources: heterogeneity among participants within a site, and heterogeneity among sites in site-average effects. Each source of unobserved heterogeneity generates problems of interpretation that can be solved only by imposing assumptions.

In the next section of the article we provide a heuristic account of the problem of unobserved heterogeneity in a single-site study. Next, we derive the assumptions needed for multisite IV analysis as a prelude to describing our estimators. We conclude with an illustrative data analysis and a brief discussion of implications for studying multiple mediators in multisite experiments.

A HEURISTIC INTRODUCTION TO IV ANALYSIS IN A SINGLE-SITE EXPERIMENT

Let us first consider conventional IV analyses based on the assumption of homogeneous treatment effects. This can be regarded as a standard path model. Next, we incorporate heterogeneity in responses to treatment, leading us to formulate a person-specific path model.

Homogeneous Treatment Effects

If we assume that all persons benefit equally from program participation, we can represent the IV model using the standard path diagram (cf. Baron & Kenny, 1986; Duncan, 1966; Wright, 1934) displayed in Figure 1: The randomly assigned treatment assignment T has a causal effect γ on program participation, conceived as the mediator, M. In turn, the impact of M on the outcome Y is δ, the effect of program participation on the outcome. A key feature of this model is that we assume the direct path between T and Y, given M, is null. This assumption is known as the "exclusion restriction" in IV analysis: T can affect Y only through its effect on M.

Using conventional path analysis, the researcher would simply regress the outcome Y on M to obtain an estimate of δ and regress M on T to obtain an estimate of γ. The product of these two estimates is regarded as "the indirect effect of T on Y through the mediator M." If the direct effect is truly null, as assumed under the exclusion restriction, the "total effect" of T on Y, which we denote as β, would be equal to the indirect effect, that is,

$$\beta = \gamma\delta \tag{1}$$

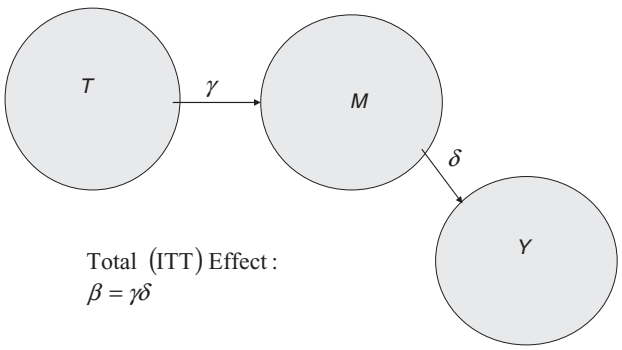

Figure 1. Instrumental variable model: Homogeneous treatment effects.

However, the regression of Y on M may yield a biased estimate of δ if persons assigned to the program will be nonrandomly allocated into levels of program participation, M. For example, students or their parents or their teachers might use various sources of information that the researcher cannot observe to guide the decision about participating in an educational program, and if decisions to participate were correlated with the outcome, the regression of Y on M would be biased.

In contrast to the conventional path analysis, the IV method for estimating δ exploits the fact that participants are randomly assigned to T. Thus, we can obtain unbiased estimates of β and γ by regressing Y on T and M on T, respectively. If the sample size is sufficiently large, the ratio of these two estimates will be a reasonable estimate of $\delta = \beta/\gamma$ so long as γ is not too close to zero.

Heterogeneous Treatment Effects

Unfortunately, the simplicity of conventional IV model (Figure 1) depends on a very strong assumption: that all participants respond the same way to treatment assignment. Although this assumption is frequently invoked without discussion in reports of IV analyses, there is good reason to expect in many if not all applications that this assumption will be false. People may vary in their motivation to participate, and some may benefit more than others from participating.

To represent heterogeneity in response to treatment assignment, we can construct a person-specific path diagram as in Figure 2. Each participant has a unique causal effect of T on M. This person-specific causal effect is denoted by Γ. We refer to Γ as "compliance," as it measures the extent to which an individual's value of M changes in response to assignment to T. The population-average value of Γ is $E(\Gamma) = \gamma$. In the same spirit, the person-specific causal effect of M on Y is Δ. We refer to Δ as the "effect" as it denotes the causal quantity of interest. The average effect in the population is $E(\Delta) = \delta$. The "total effect" of T on Y for our participant is the product of the two causal effects Γ and Δ as shown in Figure 2. Defining this total effect for our participant as B, we see then that $\mathrm{B} = \Gamma\Delta$. The key problem here is that the population average effect of T on Y is no longer the product $\gamma\delta$ as in Figure 1. Rather, this average is given by

$$\begin{aligned} E(\mathrm{B}) = \beta = E(\Gamma\Delta) &= E(\Gamma)E(\Delta) + Cov(\Gamma, \Delta) \\ &= \gamma\delta + Cov(\Gamma, \Delta) \end{aligned} \tag{2}$$

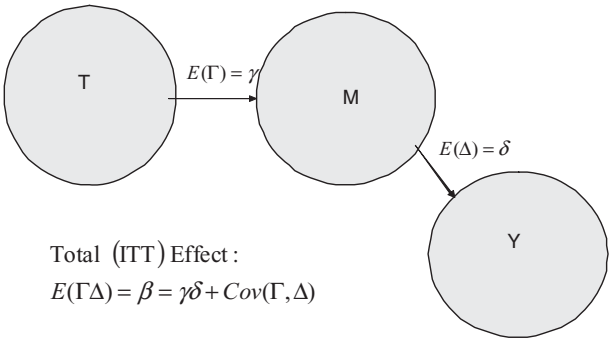

Figure 2. Instrumental variable model: Heterogeneous treatment effects.

The problem with Equation 2 is that the average intent-to-treat effect, β, depends not only on the product of the two causal effects $\gamma\delta$ as in the conventional path (Equation 1) but also on $Cov(\Gamma, \Delta)$, the covariance Γ andΔ. The presence of the term $Cov(\Gamma, \Delta)$ indicates that the population average effect of treatment assignment will be large when people who comply with the program (and thus have positive values of Γ) will tend to benefit from it (and thus have positive values of Δ). On the contrary, in a world where everyone who stood to benefit from the program refused to comply, the average impact of assignment to the program would be zero (This could happen, for example, persons who might benefit from a mental health intervention might be too depressed to attend the program.)

How can we then accomplish our aim, which is to estimate the average impact of program participation, δ when the treatment effect is heterogeneous? One option is simply to assume $Cov(\Gamma, \Delta) = 0$, there is no covariance between compliance Γ and effect Δ. In this case, Equation 2 becomes equivalent to the conventional model (Equation 1), and we can identify the average treatment effect (ATE) of M on Y as $\delta = \beta/\gamma$, $\gamma \neq 0$. However, the "no compliance-effect covariance" assumption may be implausible in many cases, particularly in cases where individuals have some knowledge of how much they will benefit from M (i.e., they know their Δ), and if that knowledge influences their compliance with assignment to T (their Γ). Thus, the no-compliance-effect covariance assumption will seem to apply either when (a) participants (or other agents such as students or physicians making assignments to M) have no foreknowledge of likely benefit from participation or (b) the benefits from participation are a constant, taking us back to the conventional model.

Rather than assuming no covariance between Γ and Δ, Angrist et al. (1996) developed an alternative approach. If T and M are binary, the authors reasoned that there are four different kinds of people. The *compliers* are those who would participate (that means select $M = 1$) if offered the program ($T = 1$) and not participate (i.e., $M = 0$) if assigned to the control ($T = 0$). Therefore, for compliers, the impact of being assigned to the novel program is $\Gamma = 1 - 0 = 1$. *Never takers* are those who would never take up the program. That means $M = 0$ regardless of treatment assignment, so for the never takers, the impact of treatment assignment is $\Gamma = 0 - 0 = 0$. *Always takers* would always take up regardless of treatment assignment, so $M = 1$ either way and for this group $\Gamma = 1 - 1 = 0$. Defiers are those who would refuse to take up M if assigned to the program (so $M = 0$ if $T = 1$) but who would participate if assigned not to (so $M = 1$ if $T = 0$). Thus, for defiers, $\Gamma = -1$.

We might assume that there are no defiers. This is also known as the *monotonicity* assumption, meaning that being assigned to the program cannot reduce the likelihood of program participation, hence $\Gamma \geq 0$ (Angrist et al., 1996). Under this assumption, we have

$$\begin{aligned} E(\mathrm{B}) = \beta &= E(\Delta\Gamma|\Gamma = 1)\Pr(\Gamma = 1) + E(\Delta\Gamma|\Gamma = 0)\Pr(\Gamma = 0) + E(\Delta\Gamma|\Gamma = -1) \\ &\quad \Pr(\Gamma = -1) \\ &= 1 * E(\Delta|\Gamma = 1)\Pr(\Gamma = 1) + 0 * E(\Delta|\Gamma = 0)\Pr(\Gamma = 0) - 1 * E \\ &\quad (\Delta|\Gamma = -1)\Pr(\Gamma = -1) \\ &= E(\Delta|\Gamma = 1)\Pr(\Gamma = 1) \\ &\equiv \delta_{late}\gamma. \end{aligned} \tag{3}$$

The last step follows from monotonicity, $\Pr(\Gamma = -1) = 0$, and from that fact that, under our assumptions γ, the average impact of assignment to the program, is the fraction of the population who comply.

So δ_{late} is the impact of program participation for the compliers – those whose participation in M is induced by assignment to $T = 1$ rather than $T = 0$. A problem for interpretation is that the magnitude of δ_{late} may depend on how effective the program is in inducing participation. A program director who is very skilled at encouraging participation in the program in one study may generate a different δ_{late} than will a program director in another study who is less skilled at doing so, even if the population average impact, ATE, is the same in the two studies.

In sum, if the benefits of program participation vary among participants (as in Figure 2), the population average "total effect" of being assigned to the treatment is no longer a simple product $\gamma\delta$, unless we invoke an assumption that is open to challenge. We have to "choose our poison:" we can invoke the no-compliance-effect covariance assumption, in which case our population parameter of interest becomes $\delta = \delta_{ATE}$, the average treatment effect in the population; or, at least in the case of binary M, we can invoke the monotonicity assumption, in which case $\delta = \delta_{LATE}$, the local average treatment effect (LATE). The monotonicity assumption is generally weaker than is the no-covariance assumption, but the gift conferred by the monotonicity assumption, namely, δ_{LATE}, carries with it interpretational difficulties that must be discussed in the context of any application.

First, LATE is a bit obscure in that, in many studies, we don't know who the compliers are. Although we know which of the participants assigned to the program actually participated, we don't know whether those participants would have participated even if assigned to the control group—unless members of the control are somehow prohibited from participating. Thus, for example, in the national Head Start Evaluation, we know that many children assigned to Head Start did participate in Head Start. However, some of these children might have attended Head Start even if assigned to the control condition. We know that because we see that a surprisingly large fraction of the control group actually turned up in Head Start centers. In cases where members of the program group cannot participate, LATE becomes "the treatment effect on the treated" (see Little & Yau, 1998), a more meaningful estimate because "the treated" are an observable group: those assigned to the novel innovation who actually participate.

Second, however, LATE is always "instrument dependent." Suppose that in one study, only a small fraction of persons who desperately needed the program complied with their treatment assignment, whereas in a second study, a larger fraction, many of whom did not need the program, participated. Then LATE in the two studies would differ even if the two study populations had the same ATE. Thus interpretation of LATE generally requires some scientific judgment about the correlates of likely compliance and likely benefit, and this judgment is open to criticism.

FORMALIZATION OF THE CAUSAL MODEL AND ITS ASSUMPTIONS

The Random Coefficient Model in a Single-Site Study

So far, we have discussed the conceptual problems posed for IV analysis in the case of heterogeneous treatment effects. In this section, we clarify all of the five key assumptions that must be met to justify the instrumental variable (IV analysis) in the case of binary T and M in a single-site study.

1. Stable Unit Treatment Value Assumption (SUTVA): For any dependent variable, each participant possesses one and only one potential outcome under each treatment assignment.

It has become standard to define $M_i(0)$ as the level of program participation for participant i if that participant is assigned to the control condition and $M_i(1)$ as the level of program participation that participant will exhibit if assigned to the program. Similarly define $Y_i(0)$ as the outcome for participant i if that participant is assigned to the control condition and $Y_i(1)$ as the outcome that participant will exhibit if assigned to the program. Thus, for each outcome variable, each participant is assumed to have a single potential outcome under each treatment assignment. Rubin (1986) called attention to this assumption, noting that it has two implications. First, SUTVA requires that one participant's outcome does not depend on another participant's treatment assignment. In our illustrative example of the Double Dose program, we will have to worry about how the assignment of other students to Double Dose might affect the outcomes of any particular focal student. Second, SUTVA requires that there is only one version of each treatment: each participant assigned to a new program experiences the same program (Rubin, Stuart, & Zanutto, 2004). We can relax this assumption a bit by assuming that, within a given site, each participant in the program experiences the same program. Hong and Raudenbush (2006) discussed strategies for relaxing SUTVA in social settings such as classrooms and schools, but we invoke SUTVA here for simplicity in this article.

2. Exclusion restriction under the random coefficient model. Our definition of potential outcomes allows us to write a person-specific linear model for program participation.[1] The observed value of M for participant i is

$$M_i = \gamma_0 + \Gamma_i T_i + \nu_i \tag{4}$$

where $M_i(0) = \gamma_0 + \nu_i$, $M_i(1) = \gamma_0 + \Gamma_i + \nu_i$, and the person-specific causal effect of treatment assignment on program participation is $M_i(1) - M_i(0) = \Gamma_i$. We can write a similarly structured random coefficient model[2] for the observed outcome

$$Y_i = \delta_0 + \Delta_i M_i + e_i \tag{5}$$

From Equation 5 it follows that $Y_i(T_i = 0) = Y_i(0) = \delta_0 + \Delta_i M_i(0) + \nu_i$, $Y_i(T_i = 1) = Y_i(1) = \delta_0 + \Delta_i M_i(1) + e_i$ and the person-specific causal effect of treatment assignment on the outcome is

$$\begin{aligned} Y_i(1) - Y_i(0) \equiv B_i &= \Delta_i [M_i(1) - M_i(0)] \\ &= \Delta_i \Gamma_i . \end{aligned} \tag{6}$$

Equation 5 implies an *exclusion restriction* because it reflects no dependence of Y on T given M, without which Equation 6 would be invalid.

3. No compliance-effect covariance—or monotonicity. We can assert $E(\Delta_i \Gamma_i) \equiv \beta = \delta\gamma$ where $\delta = \delta_{ATE}$ if we invoke the no-compliance-effect covariance assumption (Equation 2), or $\delta = \delta_{LATE}$ if we invoke the monotonicity assumption (Equation 3).

[1]In general, person-specific linearity would seem a strong assumption, but when T is binary, as in our case, it is not.

[2]Once again, if M were continuous the person-specific linearity assumption would need to be probed, but not in our case, with binary M.

4. Ignorable assignment of T. The assumption that T is effectively randomized enables to identify our causal effects. We can estimate $E(M|T=1) - E(M|T=0)$ and $E(Y|T=1) - E(Y|T=0)$ from our sample means. These differences are equivalent to the average causal effects of T on M and of T on Y under randomization, which insures

$$\begin{aligned} E(M|T=1) - E(M|T=0) &= E(M(1)|T=1) - E(M(0)|T=0) \\ &= E(M(1)) - E(M(0)) \\ &= E[(M(1) - (M(0)] \\ &= E(\Gamma) \equiv \gamma. \end{aligned} \tag{7}$$

And, similarly, $E(Y|T=1) - E(Y|T=0) = E(B) \equiv \beta$.

5. Effectiveness of the instrument. By assuming in addition that $\gamma \neq 0$, we can identify the causal impact of program participation as we can identify the causal effects our structural model as $\beta/\gamma = \delta$.

In sum, in the case of a single-site study with binary T and binary M, when we assume causal effects to be heterogeneous across participants, we can identify the causal effect of program participation (δ_{ATE} or δ_{LATE}) under the following assumptions:

(i) SUTVA,
(ii) Exclusion restriction,
(iii) No compliance-effect covariance – or – monotonicity,
(iv) ignorable assignment of T, and
(v) effectiveness of the instrument.

We now write the structural model for the single-site case as

$$\begin{aligned} Y_i &= \delta_0 + \Delta_i M_i + e_i \\ &= \delta_0 + \delta M_i + (\Delta_i - \delta) M_i + e_i \\ &= \delta_0 + \delta M_i + \varepsilon_i \end{aligned} \tag{8}$$

where $\varepsilon_i = (\Delta_i - \delta) M_i + e_i$ is the random disturbance term. If the impact of program participation is heterogenous (so that $\Delta_i - \delta$ varies from person to person), the variance of ε_i will generally vary as well. The disturbance term for those who do not participate will be e_i, whereas the disturbance term for those who do participate will be $\varepsilon_i = \Delta_i - \delta + e_i$.

Extending the Random Coefficient Model to the Multisite Designs

We now regard the results of the previous section as being replicated in each of k sites having subscripts $s = 1, \ldots, k$. Hence the average effect of T on Y in site s is β_s, the average effect of T on M in site s is γ_s, and the average effect of M on Y in site s is δ_s. The corresponding averages taken across sites are $E(\beta_s) = \beta$, $E(\gamma_s) = \gamma$, $E(\delta_s) = \delta$. We write $\beta_s = \gamma_s \delta_s$, keeping in mind that δ_s can be interpreted as the average effect of M on Y in site s if we are willing to make the assumption of no compliance-effect covariance

(Equation 2), whereas δ_s can be interpreted as the local average treatment effect in site s if we instead assume monotonicity (Equation 3). Either way, we face the problem that the population average ITT effect is

$$\beta = E(\beta_s) = E(\gamma_s \delta_s) = \gamma\delta + Cov(\gamma_s, \delta_s). \tag{9}$$

The term $Cov(\gamma_s, \delta_s)$ is covariance between the site-specific average compliance γ_s and site-average effect of program participation, δ_s. We call this the "between-site compliance-effect covariance." Its existence complicates identification. If we add a sixth assumption

(vi) between-site compliance-effect independence[3],

which implies that $Cov(\gamma_s, \delta_s) = 0$, we can estimate the ITT effects β and γ and then identify $\delta = \beta/\gamma$, assuming the instrument is effective, on average across sites. Therefore, we add

(vii) effectiveness of the instrument on average,

that is $\gamma \neq 0$.

Summary of assumptions for the multisite case. We can now collect our assumptions to incorporate multiple sites. We have

(i) SUTVA within each site,
(ii) Exclusion restriction in each site.
(iii) No compliance-effect covariance – or – monotonicity – within each site,
(iv) ignorable assignment of T within each site,
(v) effectiveness of the instrument within each site,
(vi) independence of the site-average compliance and the site-average effect of program participation, and
(vii) effectiveness of the instrument on average.

In the next section we consider three alternative approaches to identification and estimation. We see that not all of these seven assumptions are required for any of the three options we define. This will become important in selecting a method of analysis, as we illustrate with our data.

ALTERNATIVE IDENTIFICATION AND ESTIMATION STRATEGIES

We now consider three strategies for multisite IV analysis: (a) estimate the impact of participation site by site and average the estimates, (b) estimate the average impact of participation using a single instrument, and (c) use Site × Treatment interactions as multiple instruments. These three strategies rely on somewhat different assumptions and are suitable for different data sets. To reveal how they compare, we find it useful to summarize the

[3]The assumption of independence of (γ_s, δ_s) is stronger than the assumption $Cov(\gamma_s, \delta_s) = 0$. We see why the stronger assumption is required in the next section.

data from each site in terms of the estimates of the ITT effects in each site: $\hat{\beta}_s$ (the estimated effect of T on Y); and $\hat{\gamma}_s$ (the estimated effect of T on M). We then see that $\hat{\beta}_s$ is a linear function of $\hat{\gamma}_s$. The three methods differ in how they analyze this linear relationship.

To reveal the relationship between $\hat{\beta}_s$ and $\hat{\gamma}_s$, recall that our random coefficient model for site s is

$$\begin{aligned} Y_{is} &= \delta_{0s} + \delta_s M_{is} + \varepsilon_{is} \\ &= \delta_{0s} + \delta M_{is} + (\delta_s - \delta) M_{is} + \varepsilon_{is}. \end{aligned} \tag{10}$$

Within each site, we regress Y_{is} on T_{is} yielding $\hat{\beta}_s = \sum_{i=1}^{n} (T_{is} - \bar{T}_s) Y_{is} / \sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2$ where T_{is} as the treatment assignment for participant i in site s and $\bar{T}_s$ as its site mean. Substituting the right-hand side of expression (10) for Y_{is}, we find

$$\begin{aligned} \hat{\beta}_s &= \frac{\sum_{i=1}^{n} (T_{is} - \bar{T}_s) Y_{is}}{\sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2} = \frac{\sum_{i=1}^{n} (T_{is} - \bar{T}_s)[\delta_{0s} + \delta M_{is} + (\delta_s - \delta) M_{is} + \varepsilon_{is}]}{\sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2} \\ &= \frac{\delta_{0s} \sum_{i=1}^{n} (T_{is} - \bar{T}_s) + \delta \sum_{i=1}^{n} (T_{is} - \bar{T}_s) M_{is} + (\delta_s - \delta) \sum_{i=1}^{n} (T_{is} - \bar{T}_s) M_{is} + \sum_{i=1}^{n} (T_{is} - \bar{T}_s) \varepsilon_{is}}{\sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2} \\ &= \delta \hat{\gamma}_s + (\delta_s - \delta) \hat{\gamma}_s + E_s. \end{aligned} \tag{11}$$

Here $\hat{\beta}_s$ is the ordinary least squares (OLS) estimate of the average effect of T on Y in site s, $\hat{\gamma}_s = \sum_{i=1}^{n} (T_{is} - \bar{T}_s) M_{is} / \sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2$ is the OLS estimate of the average effect of T on M in site s, and $E_s = \sum_{i=1}^{n} (T_{is} - \bar{T}_s) \varepsilon_{is} / \sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2$ is a random disturbance having zero mean and variance $Var(E_s) \equiv V_s$. If the random disturbances, ε_{is} have constant variance σ^2 within sites, $V_s = \sigma^2 / \sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2$, the usual form for the sampling variance of an OLS variance.[4] However, when the effects of program participation are heterogeneous, the appropriate form for V_s is slightly more complicated, as we show in Appendix A.

Crucially, $\hat{\beta}_s$ (Equation 11) has two components of variance. The first is the sampling variance $Var(E_s) \equiv V_s$ just described. The second component is the variance of the true effects of participation that is, the variance of δ_s across sites. For notational simplicity, we set $U_s \equiv (\delta_s - \delta)$. For now, we adopt a random effects framework and define $Var(U_s) \equiv \tau_\delta^2$ meaning that τ_δ^2 is the variance of the impact of program participation across the sites in a population of sites to which we wish to generalize. In our illustrative example, our sites are high schools. Using Equation 11, we thus have the variance components model

$$\begin{aligned} \hat{\beta}_s &= \delta \hat{\gamma}_s + U_s \hat{\gamma}_s + E_s, \\ E(U_s) = E(E_s) = 0, \quad Var(U_s) &= \tau_\delta^2, \quad Var(E_s) = V_s, \quad Cov(U_s, E_s) = 0. \end{aligned} \tag{12}$$

[4]In the case of binary T as illustrated in this article, $\sigma^2 / \sum_{i=1}^{n} (T_{is} - \bar{T}_s)^2 = \sigma^2 / [n \bar{T}_s (1 - \bar{T}_s)]$. Note that in this case $\bar{T}_s$ is the fraction of cases assigned to the program in site s.

Equation 12 clarifies the logic of our three options for estimation, to which we now turn. To clarify conceptual issues, we assume for now that the variances τ_δ^2, V_s are known. In fact, we must estimate these variances from the data at hand. We show in appendices exactly how this can be done with available software.

Option A: Compute Site-by-Site Estimates, Then Average

The first strategy has two stages. First, one simply computes $\hat{\beta}_s$ and $\hat{\gamma}_s$ within each site and then divides the former by the latter. From Equation 12 we see that this yields

$$\hat{\delta}_s = \hat{\beta}_s/\hat{\gamma}_s = \delta + U_s + E_s/\hat{\gamma}_s, \tag{13}$$

For large n per site, each $\hat{\delta}_s$ is an (approximately) unbiased estimator of δ, our central parameter of interest. Second, averaging these site-specific estimates across sites can give us an approximately unbiased estimator of δ.

Option A does not require that our data adhere to assumption (vi) "between-site compliance-effect independence," nor does it require assumption (vii) "effectiveness of the instrument on average." It does, however, require assumption (v) "effectiveness of the instrument within each site," without which $\hat{\gamma}_s$ may be near zero and the division required may cause (13) to "blow up" or simply be undefined. This can be a fatal problem when assignment to T has weak effects on program participation M in some sites or when the sample size per site is so small that estimates $\hat{\gamma}_s$ veer toward zero by chance. We find (next) that the problem of near-zero compliance arises in a few sites in or own data.

Specifically, if we simply average the site-specific estimates, a few sites we obtain

$$\tilde{\delta} = \frac{\sum_{s=1}^{k} \hat{\delta}_s}{k} = \delta + \frac{\sum_{s=1}^{k} (U_s + E_s/\hat{\gamma}_s)}{k}. \tag{14}$$

Conditional on each $\hat{\gamma}_s$, our estimator has variance

$$Var(\tilde{\delta}) = \frac{Var\left(\sum_{s=1}^{k} (U_s + E_s/\hat{\gamma}_s)\right)}{k^2} = \frac{\tau_\delta^2}{k} + \frac{+\sum_{s=1}^{k} V_s/\hat{\gamma}_s^2}{k^2}. \tag{15}$$

The key problem here regards the division by $\hat{\gamma}_s^2$ in the numerator of Equation 15. If some of the sites—or even one site—exhibits very low compliance, such that $\hat{\gamma}_s^2$ is near zero, Equation 15 will "blow up" generating incredibly wide interval estimates for our unknown parameter δ. Weighting by sample size does not solve this problem.

A much more stable estimator can be constructed as a precision-weighted average. Here the "precision" is the reciprocal of variance, so the sites generating estimates $\hat{\delta}_s$ having great uncertainty (large variance) will be weighted down. Given that each $\hat{\delta}_s$ is an (approximately) unbiased estimator of δ with variance $\tau_\delta^2 + V_s/\hat{\gamma}_s^2$, we define the weight $W_s = 1/(\tau_\delta^2 + V_s/\hat{\gamma}_s^2) = \hat{\gamma}_s^2/(\hat{\gamma}_s^2\tau_\delta^2 + V_s)$. Weighting each site's estimate by W_s will sharply reduce the influence of data from sites with low compliance. Our precision-weighted estimator (and

its sampling variance; see Appendix A) will be

$$\hat{\delta}_A = \frac{\sum_{s=1}^{k} W_s \hat{\delta}_s}{\sum_{s=1}^{k} W_s}, \qquad Var(\hat{\delta}_A|\hat{\gamma}) = \frac{1}{\sum_{s=1}^{k} W_s}. \tag{16}$$

The sampling variance of $\hat{\delta}_A$, the precision-weighed estimator in Equation 16, will tend to be much smaller than that of unweighted estimator, $\bar{\delta}$ in Equation 15.[5] However, two caveats are in order. First, we cannot implement Equation 16 without knowing the value of the variances τ_δ^2, V_s. Appendix A shows how to estimate these with available software. However, in studies with few sites τ_δ^2 may be estimated poorly, introducing random variation in the weights and undermining the efficiency of the estimator. Second, $\hat{\delta}_A$ will be biased if the weights W_s are correlated with the estimates $\hat{\delta}_s$. The bias would arise if the squared site-average compliance, γ_s^2 is associated with site-average impact, δ_s. Thus, to use the precision-weighted average $\hat{\delta}_A$ as defined in Equation 16 requires that we invoke assumption (vi) "independence of site-average compliance and site-average effect." So the major advantage of Option A is lost.

Option B: Compute Global Estimates of the ITT Effects, Then Divide

The second strategy involves pooling the data from all sites in order to obtain global estimates of the β and γ, and then dividing to obtain as $\hat{\delta} = \hat{\beta}/\hat{\gamma}$. The intuitive advantage of this strategy is that estimates of the global β and γ based on pooling the data across sites will be more precise than estimates using data from a single site at a time. Site-specific estimates used in Option A induce a "finite n sampling bias" that can be appreciable in sites with small n and/or small γ_s (see Reardon et al., 2011). Option B will ease this problem as long as the global γ is not too small and is estimated relatively precisely (as it surely will in our data with 60 sites and about 12,000 students).

Option B does not require asssumption (v) "effectiveness of the instrument within each site," and that can be a major advantage, as it is for our data (see next). It does, however require (vi) "independence of site-average compliance and site-average effect" and (vii) "effectiveness of the instrument on average."

This approach can be implemented in three very simple steps: (a) Using now-standard software for hierarchical linear models (also known as "mixed models"), estimate the global average ITT effect β and the variation in the "true" site-specific ITT effects across sites, that is $Var(\beta_s) = \tau_\beta^2$. (b) Similarly, estimate the global average compliance, γ and the variation in the "true" compliances across sites $Var(\gamma_s) = \tau_\gamma^2$. Then (c) simply divide to obtain an estimate of the average impact of program participation and its (approximate) sampling variance[6]

$$\hat{\delta}_B = \hat{\beta}/\hat{\gamma}, \qquad Var(\hat{\delta}_B) \cong Var(\hat{\beta})/\hat{\gamma}^2. \tag{17}$$

[5]If we define "precision" as the inverse of variance, note that the precision of $\hat{\delta}_A$ is $\sum_{s=1}^{k} W_s$, the sum of the precisions in the k sites. We can think of $\sum_{s=1}^{k} W_s$ the amount of information these sites contain about the unknown parameter, δ. Under certain additional assumptions (see Appendix A), Equation 16 fully exploits this information, yielding a fully efficient estimator.

[6]The approximation is good only if γ is substantially different from zero and is estimated precisely.

We supply the details in Appendix B and show that this procedure is equivalent to two-stage least squares (2SLS) with a single instrument within the framework of the hierarchical linear model.

But how, then, can one recover an estimate of the variation in the impact of program participation across sites, that is $Var(\delta_s) = \tau_\delta^2$? We can answer this question by further investigating our general model (Equation 12). Using this model, we can define the "true" ITT effect for site s as

$$\beta_s = \delta\gamma_s + U_s\gamma_s. \tag{18}$$

Again invoking assumption (vi) "independence of site-average compliance and site-average-effect," we find

$$\begin{aligned} E(\beta_s) &= \delta\gamma, \\ Var(\beta_s) \equiv \tau_\beta^2 &= Var(U_s\gamma_s) = \left(\gamma^2 + \tau_\gamma^2\right)\tau_\delta^2 + \delta^2\tau_\gamma^2. \end{aligned} \tag{19}$$

By estimating the ITT effects using a hierarchical linear model, we obtain reasonably precise estimates of γ, τ_β^2, τ_γ^2, and $\delta = \beta/\gamma$ so long as we have a sufficiently large number of sites. Thus, we can estimate the variance across sites of the impact of program participation as

$$\hat{\tau}_\delta^2 = \left(\hat{\tau}_\beta^2 - \hat{\delta}^2\hat{\tau}_\gamma^2\right)/\left(\hat{\gamma}^2 + \hat{\tau}_\gamma^2\right). \tag{20}$$

Option C: Regress the ITT Effect for the Outcome *Y* on the ITT Effect for Program Participation, *M*

A third alternative is to regard Equation 12 as a regression model, where the outcome is $\hat{\beta}_s$, the site-specific ITT effect on Y, and the predictor is $\hat{\gamma}_s$ the site-specific ITT effect on M. The idea is that the slope of the line that relates $\hat{\gamma}_s$ to $\hat{\beta}_s$ represents the impact of program participation on $\hat{\beta}_s$. Rather than using OLS regression, however, we use weighted least squares with precision weights. This procedure will weight down data from sites having low compliance and/or small sample sizes. Option C offers advantages in that we do not need to assume (v) "effectiveness of the instrument within each site," or (vii) "effectiveness of the instrument on average." However, Option C does require (vi) "independence of site-average compliance and site-average effect."

To see how Option C works, note that, under assumption (vi),

$$\begin{aligned} E(\hat{\beta}|\hat{\gamma}_s) &= E(\delta\hat{\gamma}_s + U_s\hat{\gamma}_s + E_s|\hat{\gamma}_s) \\ &= \delta\hat{\gamma}_s + \hat{\gamma}_s E(U_s|\hat{\gamma}_s) + E(E_s|\hat{\gamma}_s) \\ &= \delta\hat{\gamma}_s \qquad (\textit{under assumptions}\,(iv),\ (vi)) \\ Var(\hat{\beta}|\hat{\gamma}_s) &= \hat{\gamma}_s^2\tau_\delta^2 + V_s. \end{aligned} \tag{21}$$

We can assume $E(U_s|\hat{\gamma}_s) = E(U_s) = 0$ under (vi) and $E(E_s|\hat{\gamma}_s) = E(E_s)$ under (iv) "ignorable assignment to T." With known variances, we would then construct the weight $W_{Cs} = 1/(\hat{\gamma}_s^2\tau_\delta^2 + V_s)$ and compute a weighted regression. In reality, we must estimate the variances; Appendix C shows how to do this using 2SLS within a hierarchical linear model

using k site-by-treatment instruments. Our estimator will have the form

$$\hat{\delta}_C = \frac{\sum_{s=1}^{K} W_{Cs}\hat{\gamma}_s\hat{\beta}_s}{\sum_{s=1}^{K} W_{Cs}\hat{\gamma}_s^2} \qquad Var(\hat{\delta}_C|\gamma_s) = \frac{1}{\sum_{s=1}^{K} W_{Cs}\hat{\gamma}_s^2}. \tag{22}$$

In fact, with known variances, Equation 22 is identical to the precision-weighted estimator used in Option A (Equation 16) as we show in Appendix C. The difference is that the implementation using 2SLS facilitates more flexibility in variance estimation (Appendix C). An estimate of τ_δ^2 comes naturally as output.

Fixed Effects Estimation

So far we have adopted a random effects conception. This is appropriate when the aim is to generalize to a broad universe of sites (e.g., classrooms, schools, neighborhoods), which requires sampling a reasonably large number of sites in order to insure stable estimates of the between-site variances. However, the random effects approach is not always appropriate. For example, in the Moving to Opportunity study (Kling, Liebman, & Katz, 2007), the sites were five cities (Baltimore, Boston, Chicago, Los Angeles, and New York). It makes no sense to assume we are generalizing to a broader sample of cities; in fact, interest focuses on these five important cases. In this case, sites are more appropriately regarded as fixed rather than random. We can readily obtain inferences for fixed effects models using Options A, B, and C. The approaches are equivalent to the random effects approaches described above setting $\tau_\delta^2 = \tau_\gamma^2 = \tau_\beta^2 = 0$ in which case Equation 12 is written with $U_s = \delta_s - \delta = 0$ for every s. When this assumption is false, the inferences based on the fixed-effects specification are difficult to interpret unless the model is elaborated to include interaction effects. The estimates for Option B can be obtained using standard 2SLS with fixed site effects and a single instrument. The estimates for Option C can be obtained using 2SLS with fixed site effects using k Site $\times$ Treatment interactions as instruments (see Appendix D for details). We apply these for comparison in our illustrative example.

Table 1 summarizes the assumptions required to identify the average effect of program participation under the three options. Option A with no weighting (or weighting by sample size) enables us to avoid assumption (vi) "independence of site-average compliance and site-average effect." Option A also does not require (vii) that the instrument be effective on average. Option B enables us to avoid (v) that the instrument is effective in every site, but not (vi). Option C allows us to avoid (v) and (vii) so long as the instrument is effective in at least one site. We must keep in mind, however, that some options work better than others, depending on the number of participants per site and the number of sites, as discussed previously.

ILLUSTRATIVE EXAMPLE

During the late 1990s the Chicago Public School system introduced a program called "Algebra for All," designed to insure that all ninth graders take a college-preparatory math

Table 1. Assumptions required for each method of estimation

Assumption	Option A (No Weighting)	Option A (Precision Weighting)	Option B	Option C
(i) SUTVA within each site	Yes	Yes	Yes	Yes
(ii) Exclusion restriction	Yes	Yes	Yes	Yes
(iii) No compliance-effect covariance – or – monotonicity – within each site	Yes	Yes	Yes	Yes
(iv) Ignorable assignment of T within each site,	Yes	Yes	Yes	Yes
(v) Effectiveness of the instrument within each site	Yes	Yes	No	No[a]
(vi) Independence of the site-average compliance and the site-average effect of program participation.	No	Yes	Yes	Yes
(vii) Effectiveness of the instrument on average.	No	No	Yes	No

[a]For Option c, the instrument must be effective in at least one site.

course. The results were disappointing, and it seemed that many students were not prepared for algebra (Allensworth, Nomi, Montgomery, and Lee, 2010). To solve this problem, the CPS implemented a new program in 2003: All students scoring below a cut point on the eighth-grade math test would be required to take two periods of algebra in ninth grade. Specifically, the district required students scoring the below the national median on the Iowa Test of Basic Skills (ITBS) in math to take a regular algebra class and, in addition, an extra class providing extra support in math. Students scoring above the national median were to continue to take regular algebra as before, with no extra math support. Nomi and Allensworth (2009) (henceforth "NA") used a regression discontinuity design (RDD) to estimate the impact of taking Double-Dose Algebra. Compliance with these assignments was high but not perfect, as some students scoring below the cut point did not take double dose whereas some scoring above did take double dose. Controlling for the eighth-grade math test, NA used an indicator ($T = 1$ if a student scored below the cut point, $T = 0$ if a student scored above) as an instrumental variable to estimate the impact of taking double dose ($M = 1$ if a student took double dose, $M = 0$ if not) on ninth-grade algebra achievement, Y. NA used a two-stage least squares procedure with school fixed effects. This is equivalent to our "Option B" with fixed effects, that is, with $\tau_\delta^2 = 0$. We expand on their results by analyzing the data using Options A, B, and C above with random coefficients and fixed effects models.

Data

Our data consist of first-time ninth graders who entered a regular Chicago high school in 2003. We exclude magnet schools because almost all students in magnet schools scored

Table 2. Descriptive statistics

Variables	*M*	*SD*
Scoring below the cutoff	0.40	0.49
Taking double-dose algebra	0.35	0.48
ITBS math percentile	55.49	22.09
Algebra scores	6.53	2.35

Note. ITBS = Iowa Test of Basic Skills.

above the 50th percentile on the ITBS, and these schools did not offer double-dose algebra. Also, students with identified disabilities are excluded because their curricular options differed from those of students without disabilities. The resulting sample size is 12,916 students in 60 high schools. Table 2 shows descriptive statistics. Approximately 85% of the students were eligible for free or reduced lunch; 54% were African American, 34% Latino, and 9% White.

Among students without disabilities, 40% of students scored below the 50th percentile on the eighth-grade ITBS in mathematics and 35% of students took double-dose algebra in 2003. Student achievement scores come from the algebra subset of the math portion of the PLAN examination, a standardized test developed by American College Testing (ACT Inc.). The mean of algebra scores is 6.53 with the standard deviation of 2.35.

ITT Effects

Nomi and Allensworth (2009) studied the functional form of the relationship between the eighth-grade test, X, and the ninth-grade algebra outcome Y using data from pre-policy cohorts, that is, students attending the same 60 schools prior to the treatment year, 2003. They found that a quadratic function worked well. However, the linear coefficient varied significantly from school to school. As a check on the validity of this specification, NA used prepolicy cohorts to test the association between scoring below the 2003 cut point and algebra outcome controlling for this quadratic model. Any evidence of an "effect" of scoring below the cut point during these prepolicy years would indicate model misspecification. The authors found no hint of an effect, and we have adopted their functional form here.

ITT Effect on M. We estimated a two-level hierarchical linear model with M, the indicator for taking double dose, as the outcome. Within each school, the outcome depended on a quadratic function of X = the eighth-grade math test and T, the indicator for scoring below the cut point. To ensure that between-site unobservables could not introduce confounding, we centered all student-level covariates around the school mean. When all coefficients are fixed except the intercept, this specification is equivalent to a school fixed effects model (Raudenbush, 2009). Thus, our level-1 model was

$$M_{is} = \gamma_{0s} + \gamma_s(T_{is} - \bar{T}_s) + \psi_{1s}(X_{is} - \bar{X}_s) + \psi_{2s}\left[(X_{is} - \bar{X}_s)^2 - S^2_{xs}\right] + \nu_{is} \qquad (23)$$

where γ_{0s} is a school-specific intercept, $X_{is} - \bar{X}_s$ is the eighth-grade test score X_{is} centered around the school mean $\bar{X}_s$; $(X_{is} - \bar{X}_s)^2 - S^2_{xs}$ is the square of the centered eighth-grade test score centered about its mean (that is, the school-specific variance, S^2_{xs}). At

Table 3. Model estimates for (a) ITT effect on M, (b) ITT effect on Y, and (c) effect of M on Y using Option C

	Outcome = M, ITT Effect on M	Outcome = Y, ITT Effect on Y	Outcome = Y, Effect of M on Y (Option C)
Predictor	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	0.411 (0.026)	6.301 (0.119)	6.302 (0.119)
Treatment Assignment, $T_{is} - \bar{T}_s$	0.724 (0.029)	0.354 (0.072)	
Predicted Program Participation, $\hat{M}_{is} - \bar{M}_s$			0.472 (0.094)
Pretest (linear) $X_{is} - \bar{X}_s$	$-1.42^{*}10^{-3}$ ($0.45^{*}10^{-3}$)	0.059 (0.033)	0.058 (0.003)
Pre-test (quadratic) $(X_{is} - \bar{X}_s)^2 - S^2_{\bar{X}_s}$	$-0.16^{*}10^{-4}$ ($0.17^{*}10^{-4}$)	$5.65^{*}10^{-4}$ ($0.49^{*}10^{-4}$)	$5.72^{*}10^{-4}$ ($0.51^{*}10^{-4}$)
Variances			
Variance within schools	0.067	3.053	1.748
Variance of intercept	0.039	0.846	0.846
Variance of impact	0.042	0.159	0.281
Variance of linear coefficient	$0.1^{*}10^{-4}$	$0.61^{*}10^{-3}$	$0.58^{*}10^{-3}$

level 2, we allowed γ_{0s}, γ_s, ψ_{1s} to vary and covary around their citywide means γ_0, γ, ψ_1. Of central interest were the mean γ and variance τ^2_γ of γ_s, the school-specific ITT effect of T on M.

We estimated $\hat{\gamma} = 0.724$, $SE_{\hat{\gamma}} = 0.029$, $t = 25.6$, $p < .001$ and $\hat{\tau}^2_\gamma = 0.042$, $\chi^2(59) = 552.7$, $p < .001$ (see Table 3, column 1, for details). This indicates quite strong compliance: scoring below the cut point boosted the probability of taking double dose by about 0.72 on average. The results suggest that compliance varied from school to school: Under our model, compliance in a school 1 *SD* below the mean would be about $0.724 - \sqrt{0.042} = 0.52$.

ITT Effect on Y. We estimated a model parallel to Equation 23, that is, $Y_{is} = \beta_{0s} + \beta_s(T_{is} - \bar{T}_s) + \phi_{1s}(X_{is} - \bar{X}_s) + \phi_{2s}[(X_{is} - \bar{X}_s)^2 - S^2_{xs}] + e_{is}$ with interest now focusing on the mean β and variance τ^2_β of β_s, the school-specific ITT effect of T on Y (Table 3, column 2). We estimated $\hat{\beta} = 0.354$, $SE_{\hat{\beta}} = 0.072$, $t = 4.92$, $p < .001$, and $\hat{\tau}^2_{\hat{\beta}} = 0.159$, $\chi^2(59) = 89.1$, $p = .007$. The point estimate of the average ITT effect on Y is equivalent to a standardized mean difference of $0.354/2.35 = 0.15$, and this closely corroborates the findings of NA. Note however, that this effect varies from school to school; schools with values of β_s 1 *SD*, below or above the mean would produce standardized effect sizes of about $(0.354 \pm \sqrt{0.159})/2.35 = (-0.019, 0.32)$. To the extent this variation in the ITT effect of T on Y is "explained" by variation in the ITT effect of M on Y, our models will tend to indicate that the effect of program participation (the effect of M on Y) is large, subject, of course to the assumptions required when using Options A, B, and C. The interpretation of such effects also depends upon assumptions.

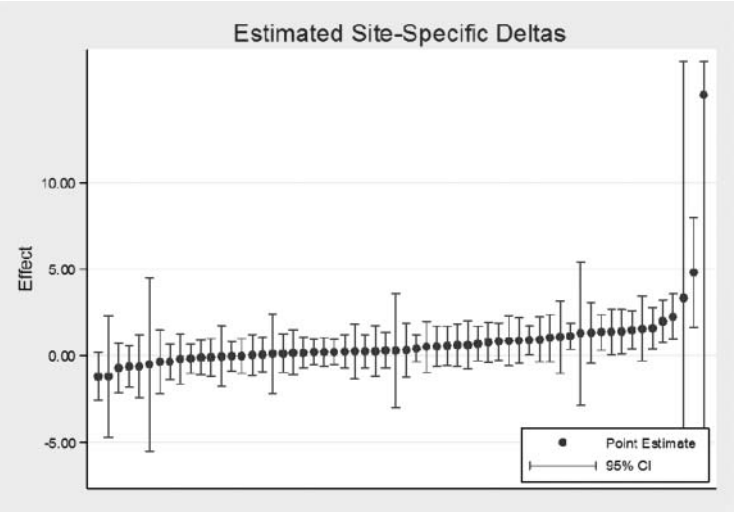

Figure 3. Site-specific estimates of program participation.

Effects of Program Participation

Option A. To compute Option A, we simply compute an OLS regression, school by school, using Equation 24, for the ITT effect on M, a regression of the same form for the ITT effect on Y.[7] We then divided the school- specific estimate of the ITT effect on Y by the school-specific estimate of the ITT effect on M (Equation 13).

It would seem that Option A would give us all we need. With about 200 students per school, it would appear that we ought to be able to obtain a reasonable estimate of the δ_s, the effect of program participation for each school s. The mean of these is an estimate of the average effect and variance of the 60 estimates would presumably characterize how much these effects vary. Alas, Figure 3 tells us why these hopes are misplaced.

The figure gives each estimate of program participation, call it $\hat{\delta}_s$, the estimated effect of program participation for sites $s = 1, \ldots, 60$, based on Equation 1. Recall that each is an estimate of either the ATE or the LATE for each site depending on our assumptions (we simply refer to the estimate as the "the estimated effect of program participation" for now.) The estimates are rank-ordered from low to high, with a 95% confidence interval around each estimate. The problem is that the confidence intervals are wide, and some are really huge (several of the confidence intervals are actually truncated to make the figure readable). For example, School 16 has an estimated effect of 15.12, about 44 times larger than the estimate of the average ITT effect mentioned above. The source of this problem is not mysterious: At School 16, the compliance effect was only 0.07, meaning that scoring below the cut point had very little effect on course taking. School 16 pretty much ignored the policy of assigning double dose to low-scoring students.

[7]Because the coefficient for the squared term in X was best viewed as fixed across schools, we used the quadratic term as an offset. That is, we used OLS school by school to estimate the model $M_{is} - \hat{\psi}_2[(X_{is} - \bar{X}_s)^2 - S^2_{xs}] = \gamma_{0s} + \gamma_s(T_{is} - \bar{T}_s) + \psi_{1s}(X_{is} - \bar{X}_s) + \nu_{is}$ where $\hat{\psi}_2$ was the estimate based on the ITT analyses already discussed.

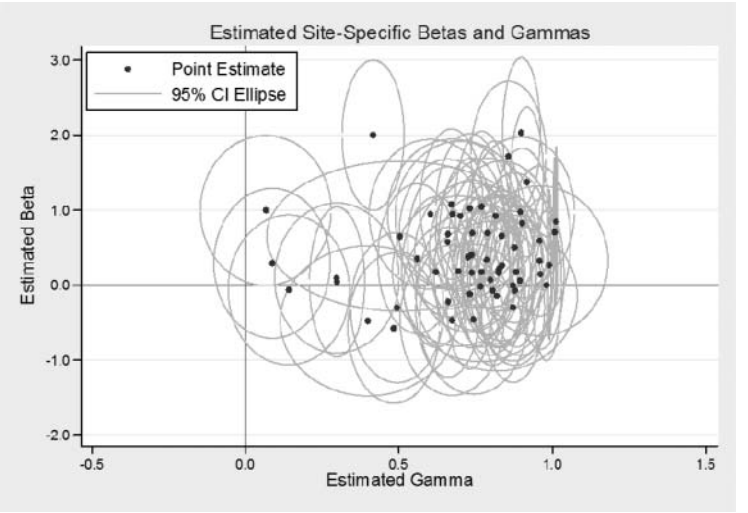

Figure 4. Plot of ITT effects on Y against ITT effects on M.

Arithmetic Average Effect. Such outliers can grossly perturb the estimate of the citywide mean. In our case, the arithmetic average effect of program participation across the 60 schools is 0.81. As we see, this point estimate is much larger than that yielded by any other method.

The implications of Figure 4 for characterizing variation across sites are worse. The sample standard deviation of the point estimates $\hat{\delta}_s$ is huge indeed at 2.12. Given that the outcome variable has a standard deviation of 2.35, this would imply that the effects of program participation are massively different across sites. But such an implication is not credible. Because the confidence intervals tend to be wide, the implication is that a huge component of variation among the values of $\hat{\delta}_s$ is simply noise. We don't want to know how much these values of $\hat{\delta}_s$ vary; rather, we want to know how much the *true values*, that is the values of δ_s vary.

Precision-Weighted Average Effect. We can address these problems by weighting. Instead of an arithmetic average, we can compute a precision-weighted average, where we define precision as the inverse of the variance of an estimate (Equation 16). Those sites producing very wide confidence intervals will receive little weight. Then some of the outliers apparent in Figure 4 will be weighted down. We accomplished this using the "V-known" (variance known) subroutine of HLM7, often used in random-effects "meta-analysis" (Raudenbush & Bryk, 1985). The level-1 model (within schools) is $\hat{\delta}_s = \hat{\beta}_s/\hat{\gamma}_s \approx \delta_s + E_s/\hat{\gamma}_s$, and we assume $Var(E_s/\hat{\gamma}_s) \approx V_s/\hat{\gamma}_s^2$ to be approximately equivalent to its true value. The level-2 model (between-schools) is $\delta_s = \delta + U_s$, where $U_s \sim (0, \tau_\delta^2)$. We obtain $\hat{\delta} = 0.547, SE_\delta = 0.100, t = 5.47, p < 0.001, \hat{\tau}_\delta^2 = 0.159, \chi^2(59) = 86.27, p = .012$. Note that the precision-weighted average of 0.547 is substantially smaller than is the arithmetic average of 0.81, and the estimated variance of 0.159 is at least plausible. Recall from our discussion of Option A, however, that the better behaved results based on the precision-weighted average are valid only under the assumption of assumption (vi) "independence of site-average compliance and site-average effect." This will also be true of estimates obtained by Options B and C (next). A problem with Option A (with or without

weighting) not shared by options B and C is the finite-sample bias of the school-specific IV estimates (Reardon et al., 2011).

Option B. Having estimated the ITT effects of T on M and T on Y, computation of estimates of the effect of program participation under Option B is simple indeed. Recall that we obtained estimated citywide average effect $\hat{\beta} = 0.354$ $(SE_{\hat{\beta}} = 0.072)$ and $\hat{\gamma} = 0.724$. The ratio is then our estimate of the global effect of program participation, $\hat{\beta}/\hat{\gamma} = \hat{\delta} = 0.490$, $SE_{\hat{\beta}} = 0.072/0.724 = 0.099$, $t = 4.95$, $p < .001$ a bit smaller than precision-weighted average from Option A of 0.547, $SE = 0.100$. To estimate the variance of the effects of program participation, we use Equation 20:

$$\hat{\tau}_{\delta}^2 \approx \frac{\hat{\tau}_{\beta}^2 - \hat{\delta}^2\hat{\tau}_{\gamma}^2}{\hat{\gamma}^2 + \hat{\tau}_{\gamma}^2} = \frac{0.159 - (0.490^2)(0.04217)}{(0.724^2 + 0.04217)} = 0.265,$$

a somewhat larger estimate of the variance of the effects of program participation across sites but rather similar to those obtained using Option C, to which we now turn. Option B does not give us an obviously credible test of the null hypothesis that $\tau_{\delta}^2 = 0$.

Option C represents the impact of program participation as a regression coefficient—representing a relationship between site-specific compliance γ_s and site-specific effect of treatment assignment on Y, $\hat{\beta}_s$. Figure 4 provides a scatter plot where $\hat{\gamma}_s$ is on the horizontal axis and $\hat{\beta}_s$ is on the vertical axis. If we fit that plot with a regression line (weighting each point by the precision of $\hat{\gamma}$) that goes through the origin (the exclusion restriction implies that if $\gamma_s = 0$ then $\beta_s = 0$), the slope of that line is an estimate of the average effect of program participation. Kling, Liebman, and Katz (2007) used a closely related approach. The variation in the "true" values of β_s around that regression line provides some evidence of variation of the site-specific effects δ_s. This is reflected in Equation 22, a regression model for $\hat{\beta}_s$ as outcome with $\hat{\gamma}_s$ as predictor and with precision weights $W_{Cs} = 1/(\hat{\gamma}_s^2\tau_{\delta}^2 + V_s)$ that depend not only on $\hat{\gamma}_s$ but also on the sampling variance V_s of each $\hat{\beta}_s$ as well as the true variance τ_{β}^2 of β_s. Equation 22 works fine as long as one knows the values of these variances (or has a good estimate). However, to obtain these variance estimates and Option C results simultaneously, we can estimate a two-level hierarchical linear model using two-stage least squares.

The first stage is simply Equation 23, which regresses program participation M on T, controlling for the pretest. These model yield, for each site, $\hat{M}_{is} = \hat{\gamma}_{0s} + \hat{\gamma}_s(T_{is} - \bar{T}_s) + \hat{\psi}_{1s}(X_{is} - \bar{X}_s) + \hat{\psi}_{2s}[(X_{is} - \bar{X}_s)^2 - S_{xs}^2]$. The second stage is a two-level hierarchical linear model $Y_{is} = \delta_{0s} + \delta_s(\hat{M}_{is} - \bar{M}_s) + \omega_{1s}(X_{is} - \bar{X}_s) + \omega_{2s}[(X_{is} - \bar{X}_s)^2 - S_{xs}^2] + r_{is}$ where $\delta_{0s}, \delta_s, \omega_{1s}$ vary randomly over schools while ω_{2s} is fixed. Appendix C describes a procedure for correcting the variance estimates, though these corrections were trivial in our data. Our result (Table 3, column 3) is $\hat{\delta} = 0.472$, $SE_{\delta} = 0.094$, $t = 5.02$ and $\tau_{\hat{\delta}}^2 = 0.281$, $\chi^2(59) = 83.71$, $p = .019$.

Fixed Effects Estimates

We also computed estimates under Options A, B, and C under a fixed effects specification. These are identical in form to those above save that all causal effects are assumed invariant across clusters, that is, $\tau_{\gamma}^2 = \tau_{\beta}^2 = \tau_{\delta}^2 = 0$. Thus, Option A uses precision weighting with weights equal to the reciprocal $\hat{\gamma}_s^2/V_s$ of the ITT effect for site s. Table 4 collects inferences

Table 4. Summary of inferences for the average effect of program participation

	Random Coefficient Model	Fixed Coefficient Model
Option A	0.547 (0.100)	0.554 (0.086)
Option B	0.490 (0.093)	0.545 (0.086)
Option C	0.472 (0.094)	0.529 (0.083)

for the average effect δ using Options A, B, and C under the random coefficient and fixed effects specifications. We note that the estimates for Options B and C based on the random coefficients model are about 11 to 12% smaller than the corresponding estimates under the fixed effects specification. One possible explanation for this result arises from the fact that the two approaches can be regarded as giving different weights to the data produced by each site. The weights diverge when the data indicate the sites to have heterogeneous effects of program participation. More investigation into these discrepancies is warranted.

DISCUSSION

A key aim of this article is to clarify the assumptions required for valid inference when using instrumental variables to estimate the effect of program participation in multisite trials. We have used a model with person-specific random coefficients to derive these assumptions. We also proposed three approaches for estimating the average effect of program participation and the variance of those effects across sites. As we have discussed (see Table 1), the three methods require somewhat different assumptions for identification, and we have seen that the statistical properties of the three methods vary as a function of the number of participants per site and the number of sites. In our example, we found quite a substantial average effect of program participation. For example, Option B with random coefficients produced a point estimate of $\hat{\delta} = 0.490$, equivalent to a standardized effect size of .21. We also found evidence of substantial heterogeneity, $\hat{\tau}_\delta^2 = 0.265$. This implies that two schools 1 *SD* below and 1 *SD* above the average would produce standardized effect sizes of $.21 \pm \sqrt{0.265} = -.01, 0.43$, respectively.

How can we interpret this finding of heterogeneity of site-average program effects? We see four possible explanations.

1. Heterogeneity of subpopulations. First, it may be that some subsets of the target population benefit more than others from participating in the program. If this were the case, and if sites vary by composition, such participant-level heterogeneity would give rise to heterogeneity in the site-average effects. For example if members of one ethnic group benefit more than do members of another ethnic group, and if sites are to some degree segregated by ethnicity, we would expect to see variation in the mean effect of participation across sites. This kind of site-level variation can, in principle, be explained by stratifying participants within sites on the participant-level characteristics that moderate the treatment effect.
2. Heterogeneity of program functioning. Second, it may be that the quality of the program differs across sites. In our example, it may be that teachers at some schools are better at teaching double-dose algebra than are teachers at other schools. Some sites may have more resources—smaller class sizes, more responsive parents—that make the

double-dose program more effective than in other sites. Peer composition might vary across sites in ways that make it easier or harder to implement the program effectively. This consideration is related to the first point just listed, but it is different. It may be that if a large fraction of one's peers are highly responsive to the program, an intellectual climate emerges that would help even those who are not part of this more responsive group.

3. Treatment by Site interactions. It may be that even if the participants in each site were similar, and even if the program is equally well implemented, that site characteristics would magnify the impact of the program. In the case of double dose, for example, if some schools are particularly good at teaching science to low-achieving students, and if the science lessons help develop those students' algebra skills, then students taking double dose may benefit more in those schools.
4. The sites may vary in terms of the size and composition of the subset of students who comply with their treatment assignment. Suppose, for example, that teachers at some sites know which students stand to gain most from double dose and that those teachers are particularly skilled at convincing those students to take double dose. In contrast, imagine that teachers at other schools are either less knowledgeable about who will benefit or less skilled in convincing those students to participate. The two sets of schools would then likely have different sets of compliers even when the two sites have the similar student composition. LATE (the treatment effect on the compliers) would then vary even if ATE (the average treatment effect in the population) were constant across sites. We note that although compliance was quite high on average in our study, it was much larger in some sites than others. There are several explanations for low compliance. In some schools, it could be that many of those scoring below the cut point did not take double dose. In other schools, many students scoring above the cut point nevertheless did take double dose. These different kinds of noncompliance imply that the subset of compliers does indeed vary across schools, and this variation could generate heterogeneity in site-average effects of program participation even if ATE is the same in all sites. Recall that we can define the average effect of program participation in a site as ATE if we are willing to assume that, within sites, there is no covariance between compliance and the effect of program participation. Alternatively, we can define the average effect of program participation in a site as LATE if we are unwilling to make this no covariance assumption, but we are willing to assume there are no defiers. Unfortunately, neither one of these choices solves the problem of heterogeneity across sites generated by the composition of compliers.

This discussion implies that considerably more inquiry is needed into ways of probing heterogeneity in site-average treatment effects. Heckman and Vytlacil (2005) created a framework for conceptualizing heterogeneity using the idea of a marginal treatment effect, and this may prove productive in clarifying sources of heterogeneity.

Multiple Mediators

We are concerned about the exclusion restriction and SUTVA in our case because the Double-Dose policy changes not only student course taking but also classroom composition. We have found that students scoring below the cut point tend to have lower achieving peers in their classrooms than do students scoring above the cut point. If peer composition affects

the outcomes, the interpretation of the impact of program participation is unclear. We intend to investigate whether peer composition helps us understand the apparently negative effect of program participation in some sites. This implies the need for a model with multiple mediators using methods akin to those of Kling et al. (2007). We might use course-taking and peer composition as two mediators of the impact of scoring below the cut point. How does our framework incorporate this possibility? Options A and B will no longer be viable in this case. One cannot estimate the effects of two mediators given a single instrument, either site by site as in Option A or globally via Option B. Option C then becomes the only viable solution. Reardon and Raudenbush (2011) derived the required assumptions for using site-specific instruments to identify the impacts of multiple mediators. They show that additional assumptions are needed beyond those listed in Table 1. More work is needed on how to evaluate these assumptions and the sensitivity of inferences to failure of the assumptions.

ACKNOWLEDGMENTS

The work reported here has been supported by the William T. Grant Foundation under the grant "Building Capacity for Evaluating Group-Level Interventions" and by the Institute of Education Sciences under the grant "Using Instrumental Variables Analysis Coupled with Rigorous Multi-Site Impact Studies to Study the Causal Paths by which Educational Interventions Affect Student Outcomes" (grant R305D090009).

REFERENCES

Administration for Children, Youth, and Families (2010). *Head Start Impact Study final report*. Washington, DC: Author.

Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (2010). College preparatory curriculum for all: Academic consequences of requiring Algebra and English I for ninth graders in Chicago. *Education Evaluation and Policy Analysis*, *31*, 367–391.

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

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2012). *Estimating Variation in Program Impacts: Theory, Practice and Applications,* New York: MDRC.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review 8*(2), 225–246.

Dempster, A. P., Laird, N. M., Rubin, & D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, *39*(1), 1–38.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, *72*, 1–16.

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

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of an intervention. *Journal of Econometrics*, *30*, 239–267.

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

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*, 669–738.

Hedges, L. V., & Olkin, I. O. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of casual inference for multi-level observational data. *Journal of American Statistical Association 101*(474), 901–910.

Kirk, R. E. (1982). *Experimental Design*. Belmont CA: Cole.

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

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

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *The Economic Journal*, *111*, 1–28.

Little, R. J., & Yau, H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods 3*(2), 147–159.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects, *Biometrika*, *74*(4), 817–827.

Nomi, T., & Allensworth, E. (2009). "Double-dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*, 111–148.

Nye, B., Konstantopoulos, S., and Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis 26*(3), 237–257.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Journal of Education, Finance and Policy 4*(4), 468–491.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*(2) 75–98.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models (Second Edition*). Thousand Oaks, CA: Sage Publications.

Reardon, S. F., & Raudenbush, S. W. (2011). *Under what assumptions do site-by-treatment instruments identify average causal effects?* Unpublished manuscript, Stanford University, Palo Alto, CA.

Reardon, S. F., Unlu, F., Zhu, P., & Bloom, H. (2012). *Bias and Bias Correction in Multi-Site Instrumental Variables Analysis of Heterogeneous Mediator Effects*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness. March 8–10, 2012: Washington, DC.

Rubin, D. B. (1986). Statistics and causal inference: Comment: which ifs have causal answers. *Journal of the American Statistical Association 81*(396), 961–962.

Rubin, D. B., Stuart, E. A., & Zaunutto, E. L. (2004). A potential outcomes view of value added assessments in education. *Journal of Educational and Behavioral Statistics*, *29*, 103–116.

Shin, Y., & Raudenbush, S. W. (2011). The causal effect of class size on academic achievement: Multivariate instrumental variable estimators with data missing at random. *Journal of Educational and Behavioral Statistics*, *36*, 154–185.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*, 298–318.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215.

APPENDICES

These appendices provide analytic details for the methods described in this article. Appendices A, B, and C consider analyses for multisite randomized trials. Appendix D considers the fixed effects case. Appendix E considers the RDD design within each site.

APPENDIX A

Site-by-Site ESTIMATION for Option A (Randomized Trials)

This is a two-step procedure. One first obtains a point estimate and standard error of the impact of program participation in site s. These provide input into a second step, the between-site analysis, yielding an estimate of the average impact and the variance of the impacts across sites.

Step 1. The first step is to compute $\hat{\delta}_s$ site by site. This can be accomplished by applying the standard 2SLS approach to Equation 10. Standard packages will estimate a first-stage equation $M_{is} = \gamma_{0s} + \gamma_s T_{is} + v_{is}$ using OLS, then substitute $\hat{M}_{is} = \hat{\gamma}_{0s} + \hat{\gamma}_s T_{is}$ for M_{is} in Equation 11, yielding the stage 2 model $Y_{is} = \delta_{0s} + \delta_s \hat{M}_{is} + \varepsilon_{is}$. Applying OLS to this equation gives (approximately) unbiased estimates $\hat{\delta}_{0s}, \hat{\delta}_s$. When computing the standard error, use M_{is} rather than $\hat{M}_{is}$ to obtain an estimate $\hat{\sigma}^2 = \sum_{i=1}^{n} (Y_{is} - \hat{\delta}_{0s} - \hat{\delta}_s M_{is})^2/(n - 2)$ of the variance.

This approach assumes homogeneity of treatment effect within sites. To reflect heterogeneity, we need to estimate at least 2 separate variances. Note that in Equation 11, $\varepsilon_{is} = e_{is} + M_{is}(\Delta_{is} - \delta_s)$, so that $Var(\varepsilon_{is}|M_{is} = 0) = Var(e_{is}) \equiv \sigma_0^2$; while $Var(\varepsilon_{is}|M_{is} = 1) = Var[e_{is} + (\Delta_{is} - \delta_s)] \equiv \sigma_0^2 + \sigma_\Delta^2 + 2\sigma_{0\Delta} \equiv \sigma_1^2$. So the variance for those who participate depends on the variance of the impact of participating, defined here as σ_Δ^2 as well as $\sigma_{0\Delta} \equiv Cov(e_{is}, \Delta_{is})$. We could elaborate further to reflect heterogeneity of Γ_{is}, implicit in the first-stage equation. A practical problem emerges when sample sizes are small or modest: Estimating multiple variances to reflect heterogeneity may backfire if these estimates are unstable, inserting extra instability into the weights we now examine. In our experience, estimating a single variance has worked well, but this may not always be true.

Step 2. The output from Step 1 is, for each site, a point estimate $\hat{\delta}_s$ and an estimated standard error, $S_{\hat{\delta}_s}$. At Stage 2, compute the estimated sampling variance $V_s = S_{\hat{\delta}_s}^2$ and assume it to be the true sampling variance. Our problem is now equivalent to a "meta-analysis" and we apply the results of Raudenbush and Bryk (1985) to this problem in the framework of a hierarchical linear model. The level-1 Equation is

$$\hat{\delta}_s = \delta_s + E_s, \quad E_s \sim (0, V_s). \tag{A1}$$

Here we regard our $\hat{\delta}_s$ as an unbiased estimate of the true effect δ_s having a known sampling variance $V_{\hat{\delta}s}$. The assumption that our squared standard error estimate is equivalent to the true sampling variance implies that this approach requires a reasonably large sample in each site as well as compliance γ_s not too far from zero. However, we will weight down data from sites with low compliance, offsetting the latter concern. The level-2 Equation is

$$\delta_s = \delta + U_s, \quad U_s \sim \left(0, \tau_\delta^2\right). \tag{A2}$$

Thus, the true effects δ_s vary over sites with variance τ_δ^2. The mixed model is

$$\hat{\delta}_s = \delta + U_s + E_s E(\hat{\delta}_s) \approx \delta_s Var(\hat{\delta}_s) = \tau_\delta^2 + V_{\delta s} \equiv W_s^{-1}. \quad \text{(A3)}$$

The parameters δ, τ_δ^2 can be estimated efficiently by now-conventional packages for hierarchical aka "mixed" linear models. We used HLM 7.0 with the "V-known" option. The program begins with simple estimates

$$\hat{\delta}^{(0)} = \sum_{s=1}^{k} \hat{\delta}_s / k; \quad \hat{\tau}_\delta^{2(0)} = \sum_{s=1}^{k} (\hat{\delta}_s - \hat{\delta}^{(0)})^2/(k-1) - \sum_{s=1}^{k} V_s/k. \quad \text{(A3)}$$

Using an EM algorithm (Dempster, Laird, & Rubin, 1977) we obtain, at convergence

$$\hat{\delta}^{(\infty)} = \sum_{s=1}^{k} W_s^{(\infty)} \hat{\delta}_s / \sum_{s=1}^{k} W_s^{(\infty)}{}_s \hat{\delta}_s, \quad Var(\hat{\delta}^{(\infty)}) \approx \left(\sum_{s=1}^{k} W_s^{(\infty)} \right)^{-1} \quad \text{(A4)}$$

where $W_s^{(\infty)} = (\tau_\delta^{2(\infty)} + V_s)^{-1}$ and the superscripts indicate the iteration number. These are maximum likelihood estimates under normality but can be regarded as method-of-moments estimates via weighted least squares. (See Raudenbush and Bryk, 2002, Chap. 14 for details.) They are maximum likelihood under multivariate normality of level-1 and level-2 random errors.

APPENDIX B

Global Estimation (Randomized Trials)

Option B is perhaps the simplest to implement and works well even when site sample sizes are small so long as the number of sites is large. There are three steps: (a) estimate β, the ITT effect on Y and τ_β^2, the variance of these across sites; (b) similarly, estimate γ, the ITT effect on M and τ_γ^2, the variance of these across sites; (c) combine results from these two steps to obtain estimates of δ and τ_δ^2.

Step 1. Estimate a two-level hierarchical linear model using one of many now-standard software packages. The level-1 model represents variation among participants within sites. Substituting M into our structural model (Equation 10), we have for each site the reduced form

$$\begin{aligned} Y_{is} &= \delta_{0s} + \delta_s M_{is} + \varepsilon_{is} \\ &= \delta_{0s} + \delta_s(\gamma_{0s} + \gamma_s T_{is} + \nu_{is}) + \varepsilon_{is} \\ &= \beta_{0s} + \beta_s T_{is} + b_{is}, b_{is} \sim \left(0, \sigma_b^s\right) \end{aligned} \quad \text{(B1)}$$

where $\beta_{0s} = \delta_{0s} + \delta_s \gamma_{0s}$, $\beta_s = \delta_s \gamma_s$, and $b_{is} = \delta_s \nu_{is} + \varepsilon_{is}$. The assumption of constant variance σ_b^s is not entirely plausible and we can readily estimate separate variances for the program and control group. We find it useful to modify (B1) slightly to avoid confounding between unobserved site characteristics and treatment assignment by using the site-mean centered predictor $(T_{is} - \bar{T}_s)$ in the place of T_{is}. Raudenbush (2009) shows that, when we regard the coefficients β_s to be fixed but allow the intercept β_{0s} to be random, our inferences

about β are identical to those based on a site fixed effects model. However, we shall also allow the effects β_s to vary randomly across sites. Thus, at level 2, we have

$$\begin{aligned} \beta_{0s} &= \beta_0 + U_{0s} \\ \beta_s &= \beta + U_s \end{aligned} \tag{B2}$$

and we specify $Var(U_{0s}) = \tau^2_{\beta_0}$, $Var(U_s) = \tau^2_{\beta}$, $Cov(U_{0s}, U_s) = \tau_{\beta_0\beta}$. We used HLM7.0; using a combination of EM algorithm (Dempster et al., 1977) and Fisher scoring (Longford, 1987), these provide estimates that are maximum likelihood under normal theory and approximately method of moments more generally. The key output is $\hat{\beta}$, $\hat{\tau}^2_{\beta}$ and their standard errors.

Step 2. Using an entirely analogous procedure, use a two-level analysis to obtain estimates of γ andτ^2_{γ}.

Step 3. Use results given by Equations 20 and 21 to obtain estimates of δ and τ^2_{δ}.

Equivalence to 2SLS with a single instrument. Suppose we use the global estimates of γ_0 and γ to generate predicted values of program participation using the single equation $\hat{M}_{is} = \hat{\gamma}_0 + \hat{\gamma} T_{is}$. We then substitute in Equation 11, our second stage. This will produce the same result as describe above for Option B. To show this simply, write the centered version of Equation 11:

$$Y_{is} - \bar{Y}_s = \delta_s(\hat{M}_{is} - \bar{M}_s) + \varepsilon_{is} - \bar{\varepsilon}_s. \tag{B3}$$

Our OLS estimator within each site will then yield

$$\hat{\delta}_s = \frac{\sum_{s=1}^{k} (\hat{M}_{is} - \bar{M}_s)(Y_{is} - \bar{Y}_s)}{\sum_{s=1}^{k} (\hat{M}_{is} - \bar{M}_s)^2} = \frac{\sum_{s=1}^{k} \hat{\gamma}(T_{is} - \bar{T}_s)(Y_{is} - \bar{Y}_s)}{\sum_{s=1}^{k} \hat{\gamma}^2(T_{is} - \bar{T}_s)^2} = \frac{\hat{\beta}_s}{\hat{\gamma}} \tag{B4}$$

Then modeling $\hat{\beta}_s$ as in Option B leads to the same estimator as Equation 20.

APPENDIX C

Option C: Regress $\hat{\beta}_s$ on $\hat{\gamma}_s$ Using Weighted Least Squares (Randomized Trials)

Option C regards $\hat{\beta}_s$ as an outcome and $\hat{\gamma}_s$ as a predictor. If we had prior knowledge of the variances τ^2_{δ}, V_s, we could estimate δ using Equation 22 by means of weighted least squares. Alas, we must estimate these variances. To do so, we operate in two steps: (a) use 2SLS within a hierarchical linear model with k Site $\times$ Treatment interactions serving as instruments; (b) substitute the actual M for the predicted M to obtain correct variance estimates.

Let us now use the site-specific estimates of $\hat{\gamma}_{0s}$ and $\hat{\gamma}_s$ to generate predicted values of program participation using the k prediction equations $\hat{M}_{is} = \hat{\gamma}_{0s} + \hat{\gamma}_s T_{is}$. We then substitute

in Equation 10, our second stage, note the OLS estimator within each site is

$$\hat{\delta}_s = \frac{\sum_{s=1}^{k} \hat{\gamma}_s(T_{is} - \bar{T}_s)(Y_{is} - \bar{Y}_s)}{\sum_{s=1}^{k} \hat{\gamma}_s^2(T_{is} - \bar{T}_s)^2} = \frac{\hat{\beta}_s}{\hat{\gamma}_s} = \frac{\hat{\gamma}_s\delta + \hat{\gamma}_s U_s + E_s}{\hat{\gamma}_s} = \delta + U_s + E_s/\hat{\gamma}_s. \quad \text{(C1)}$$

The optimal estimator has the same form as that in Option A, which, as we have said, is equivalent to Option C except that the two approaches use different methods to estimate the within-site variances. Our recommendation for variance estimation involves three steps:

1. Estimate the 2SLS equation $Y_{is} - \bar{Y}_s = \delta_s(\hat{M}_{is} - \bar{M}_s) + \varepsilon_{is} - \bar{\varepsilon}_s$, $\delta_s \sim N(\delta, \tau_\delta^2)$ using a hierarchical/mixed model yielding;
2. Now substitute M_{is} for $\hat{M}_{is}$, holding constant the estimate $\hat{\delta}$ from step 1 but reestimating the variances σ_ε^2, τ_δ^2.
3. If necessary, iterate: holding constant the variance estimates from step 2, reestimate the second stage model $Y_{is} - \bar{Y}_s = \delta_s(\hat{M}_{is} - \bar{M}_s) + \varepsilon_{is} - \bar{\varepsilon}_s$, $\delta_s \sim N(\delta, \tau_\delta^2)$; then reestimate the variances substituting M_{is} for $\hat{M}_{is}$ and holding constant $\hat{\delta}$.

In our experience to date, no iteration is needed. Indeed, the results hardly change from step 1 to step 2.

APPENDIX D

Fixed Effects Models

If sites are regarded as fixed rather than random, Options A, B, and C may be employed as above setting $U_s = 0$, equivalent to assuming $\tau_\delta^2 = 0$. Then the weight needed for precision weighting in Option A (Equation 26), which was $W_s = \hat{\gamma}_s^2/(\hat{\gamma}_s^2\tau_\delta^2 + V_s)$ in the random effects case becomes $W_s = \hat{\gamma}_s^2/V_s$. Option B can be implemented with 2SLS using site fixed effects and a single-stage 1 equation $\hat{M}_{is} = \hat{\gamma}_0 + \hat{\gamma}T_{is}$. To implement Option C, use 2SLS with site fixed effects and k stage 1 equations $\hat{M}_{is} = \hat{\gamma}_{0s} + \hat{\gamma}_s T_{is}$.

However, if the assumption $\tau_\delta^2 = 0$ is false, the estimated standard errors based on the fixed effects models will be incorrect and the point estimates will be inefficient (Reardon & Raudenbush, 2011). A correction for clustering is then essential. More important, the estimate of δ will be difficult to interpret because it may not characterize any site well, and with no quantification of heterogeneity, readers of the research may be misled. We therefore recommend testing the null hypothesis $H_0 : \delta_s = \delta, s = 1, \ldots, S$. The test is

$$H = \frac{\hat{\sigma}^2 \sum_{s=1}^{K} w_s\hat{\gamma}_s^2(\hat{\delta}_{As} - \hat{\delta}_A)^2}{\sum_{s=1}^{K} w_s\hat{\gamma}_s^2} \quad \text{(D1)}$$

and H will be distributed as approximately chi-square with k-1 degrees of freedom under the null hypothesis. If H_0 is rejected, one may seek to identify homogeneous subsets of sites, much as in the fixed effects meta-analysis literature (cf. Hedges & Olkin, 1985). An

alternative strategy is to formulate and test a priori hypotheses regarding the k site effects much as in the experimental design literature (cf. Kirk, 1982).

APPENDIX E

Adjustments for RDD

So far, we have assumed that participants are randomly assigned to $T = 1$ versus $T = 0$ within sites. However, in many studies, assignment may be based on a known rule, for example, the cut score on the pretest as in our case. The logic described earlier applies, but some modifications are needed in the analysis. We used a polynomial functional form to represent the association between the pretest and the probability of program participation as well as the association between the pretest and the outcome Y. The impact of $T = 1$ if scoring below the cut point versus $T = 0$ for scoring above, is then a discontinuity in this polynomial.

Option A. To estimate the ITT effects for each site, we fit the models

$$\begin{aligned} Y_{is} &= \beta_{0s} + \beta_s(T_{is} - \bar{T}_s) + \phi_{1s}(X_{is} - \bar{X}_s) + \phi_{2s}\left[(X_{is} - \bar{X}_s)^2 - S^2_{xs}\right] + \varepsilon_{is} \\ M_{is} &= \gamma_{0s} + \gamma_s(T_{is} - \bar{T}_s) + \psi_{1s}(X_{is} - \bar{X}_s) + \psi_{2s}\left[(X_{is} - \bar{X}_s)^2 - S^2_{ms}\right] + \upsilon_{is} \end{aligned} \quad \text{(E1)}$$

where β_{0s} and γ_{0s} are school-specific intercepts, X_{is} is the eighth-grade test score centered around the cutoff score (i.e., the 50th percentile) for student i in site s and is centered around the site mean, $\bar{X}_s$; $(X_{is} - \bar{X}_s)^2 - S^2_{xs}$ is the square of the centered eighth-grade test score centered about its mean (i.e., the school-specific variance, S^2_{xs}). This enabled us to compute for each site the outcome $\hat{\delta}_{AS} = \hat{\beta}_s/\hat{\gamma}_s$ and then use the methods for fixed and random coefficient modeling as described earlier.

Option B. To obtain estimates of the overall mean ITT effect β and compliance effect γ in the random effects case, we estimated 2SLS models of the form

$$\begin{aligned} Y_{is} &= \beta_{0s} + \beta(T_{is} - \bar{T}_s) + (\beta_s - \beta)(T_{is} - \bar{T}_s) + \phi_{1s}(X_{is} - \bar{X}_s) \\ &\quad + \phi_{2s}\left[(X_{is} - \bar{X}_s)^2 - S^2_{xs}\right] + e_{is} \\ M_{is} &= \gamma_{0s} + \gamma(T_{is} - \bar{T}_s) + (\gamma_s - \gamma)(T_{is} - \bar{T}_s) + \psi_{1s}(X_{is} - \bar{X}_s) \\ &\quad + \psi_{2s}\left[(X_{is} - \bar{X}_s)^2 - S^2_{xs}\right] + \upsilon_{is} \end{aligned} \quad \text{(E2)}$$

where $(\beta_s - \beta) \sim (0, \tau^2_\beta)$; $(\gamma_s - \gamma) \sim (0, \tau^2_\gamma)$. We then computed our estimate $\hat{\delta} = \hat{\beta}/\hat{\gamma}$. The fixed effects models were identical save that we set $(\beta_s - \beta) = (\gamma_s - \gamma) = 0$ for all s. This fixed coefficient specification is equivalent to a standard 2SLS approach with fixed site effects.

Option C used the first-stage equation

$$\begin{aligned} M_{is} &= \gamma_{0s} + \gamma(T_{is} - \bar{T}_s) + (\gamma_s - \gamma)(T_{is} - \bar{T}_s) + \psi_{1s}(X_{is} - \bar{X}_s) + \psi_{2s}\left[(X_{is} - \bar{X}_s)^2\right. \\ &\quad \left. - S^2_{xs}\right] + \upsilon_{is} \end{aligned} \quad \text{(E3)}$$

just as in E2 except that now $\hat{\gamma}_s$ was treated as a fixed coefficient estimated for each sites using OLS with fixed effects. The second stage equation was

$$Y_{is} = \delta_{0s} + \delta(\hat{M}_{is} - \bar{M}_s) + (\delta_s - \delta)(\hat{M}_{is} - \bar{M}_s) + \pi_{1s}(X_{is} - \bar{X}_s) + \pi_{2s}\big[(X_{is} - \bar{X}_s)^2 - S^2_{xs}\big] + \varepsilon_{is} \tag{E4}$$

where $\hat{M}_{is}$ is extracted from E3. In the random effects case, we regarded $Var(\delta_s - \delta) = \tau^2_\delta$ whereas in the fixed effects case we assumed $(\delta_s - \delta) = 0$ for all s. See Appendix C for a procedure to correct the within-site variance estimates.