NCEE 2009-4065 U. S. DEPARTMENT OF EDUCATION

Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher Practice and Student Achievement Outcomes?

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher Practice and Student Achievement Outcomes?

October 2009

Peter Z. Schochet
Mathematica Policy Research

Abstract

For RCTs of education interventions, it is often of interest to estimate associations between student and mediating teacher practice outcomes, to examine the extent to which the study's conceptual model is supported by the data, and to identify specific mediators that are most associated with student learning. This paper develops statistical power formulas for such exploratory analyses under clustered school-based RCTs using ordinary least squares (OLS) and instrumental variable (IV) estimators, and uses these formulas to conduct a simulated power analysis. The power analysis finds that for currently available mediators, the OLS approach will yield precise estimates of associations between teacher practice measures and student test score gains only if the sample contains about 150 to 200 study schools. The IV approach, which can adjust for potential omitted variable and simultaneity biases, has very little statistical power for mediator analyses. For typical RCT evaluations, these results may have design implications for the scope of the data collection effort for obtaining costly teacher practice mediators.

NCEE 2009-4065
U.S. DEPARTMENT OF EDUCATION

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to develop methods for assessing mediational analyses in education evaluations. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
John Q. Easton
Acting Commissioner

October 2009

Schochet, Peter Z. (2009). *Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher Practice and Student Achievement Outcomes?* (NCEE 2009-4065). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The author for this report, Dr. Peter Z. Schochet, is an employee of Mathematica Policy Research with whom IES contracted to develop the methods that are presented in this report. Dr. Schochet and other MPR staff do not have financial interests that could be affected by the content in this report.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

In support of this mission, NCEE promotes methodological advancement in the field of education evaluation through investigations involving analyses using existing data sets and explorations of applications of new technical methods, including cost-effectiveness of alternative evaluation strategies. The results of these methodological investigations are published as commissioned, peer reviewed papers, under the series title, Technical Methods Reports, posted on the NCEE website at http://ies.ed.gov/ncee/pubs/. These reports are specifically designed for use by researchers, methodologists, and evaluation specialists. The reports address current methodological questions and offer guidance to resolving or advancing the application of high-quality evaluation methods in varying educational contexts.

This NCEE Technical Methods paper addresses whether typical large-scale RCT designs have sufficient statistical power for meditational analyses that associate teacher practice and student achievement outcomes. These exploratory analyses are important for helping to understand key pathways through which the intervention affects student learning as hypothesized by the study’s conceptual model, and for identifying specific teacher practices that are most associated with student learning. These analyses, however, will be informative only if the study has sufficient statistical power for estimating mediator-achievement associations that are likely to be observed in practice; if not, there will be a low chance of finding statistically significant associations. This power issue is critical for designing education RCTs due to the high cost of obtaining teacher practice data through classroom observations and videotaping. The main conclusion from this paper is that for typical RCTs with 60 schools, statistical power is likely to be limited for associating teacher practice and student achievement outcomes using ordinary least squares (OLS) methods, and especially using instrumental variable (IV) methods.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

Randomized control trials (RCTs) in the education field often test interventions that aim to improve teacher practices, with the ultimate goal of increasing student academic achievement. These interventions typically provide enhanced services to teachers, such as training in a new reading or math curriculum, mentoring services, or the introduction of new technologies or materials in the classroom. Consequently, the conceptual model for these RCTs posits that improvements in student outcomes are mediated by treatment-induced improvements in teacher practices.

Given this conceptual model, RCTs often collect data on mediating teacher practice outcomes (using classroom observation protocols, videotaping, principal ratings, and teacher logs or surveys) and on student outcomes (such as achievement test scores). These data are then typically used to estimate impacts (mean treatment-control differences) on both sets of outcomes.

For these RCTs, there is also often interest in conducting analyses to link the impact estimates on the teacher practice and student outcomes (Baron and Kenny 1986; Gamse et al. 2008; Jackson et al. 2007; Holland 1988; MacKinnon and Dwyer 1993; Sobel 2008). These exploratory analyses are often conducted using regression methods to estimate the association between the two sets of outcomes. These mediator analyses aim to assess the extent to which the study's conceptual model is supported by the data, and to identify pathways—specific dimensions of teacher practices represented by the mediators and their subscales—through which the intervention improves the classroom environment and student learning.

In RCTs in the education area, sample sizes are typically selected so that the study will have sufficient power for detecting impacts on student outcomes—and in particular, on student achievement test scores—that are deemed to be educationally meaningful and attainable (for example, 0.25 standard deviations). In assessing appropriate sample sizes, some RCTs also consider power levels for detecting impacts on teacher practice outcomes. Thus, there is a growing literature in the education field on methods to calculate statistical power for detecting impacts on student outcomes (Hedges and Hedberg 2007; Raudenbush 1997; Schochet 2008) and mediating outcomes (Raudenbush et al. 2008).

There is also a large literature on methods for calculating statistical power for regression coefficients under non-clustered designs (see Cohen 1977, 1988; Kramer and Thiemann 1987; MacCallum et al. 1996; and Rogers and Hopkins 1988). However, the literature has not addressed statistical power issues for regression-based mediator analyses for the types of large-scale clustered RCT designs that are typically used in education research. These methods are needed to assess whether typical RCT samples (for example, 60 schools and 180 classrooms) have sufficient power for detecting associations between teacher practice mediators and student outcomes that are likely to hold in practice. This issue is important, because it could influence decisions about the scope of data collection for teacher practice measures, which tends to be very costly, especially if classroom observations are conducted and videotapes and observation protocols are coded for scale construction. If power levels are low for mediator analyses—that is, if there is little chance that significant mediator-test score relationships can be found—the teacher practice mediators may have limited value for the study beyond a heuristic, qualitative linking of the mediating and student outcomes (and, hence, impacts).

This report is the first to systematically examine, both theoretically and empirically, the calculation of statistical power for regression-based mediator analyses for clustered RCTs in the education area. The focus is on the most commonly-used clustered design where schools are randomly assigned to a single treatment or control condition. The report develops formulas for calculating statistical power for mediator analyses using two regression approaches: (1) a simple ordinary least squares (OLS) approach where the student outcome is regressed on a single mediator and (2) an instrumental variables (IV) approach where treatment status is used as an instrument for the mediator. The formulas also incorporate the effects of measurement error in the mediator. Finally, the report uses the developed formulas to simulate the statistical power of mediator analyses that aim to associate teacher practice and student test score outcomes. This analysis attempts to answer the key question: How many study schools are required to ensure that RCTs of education interventions have enough statistical power for linking teacher practice and student achievement outcomes?

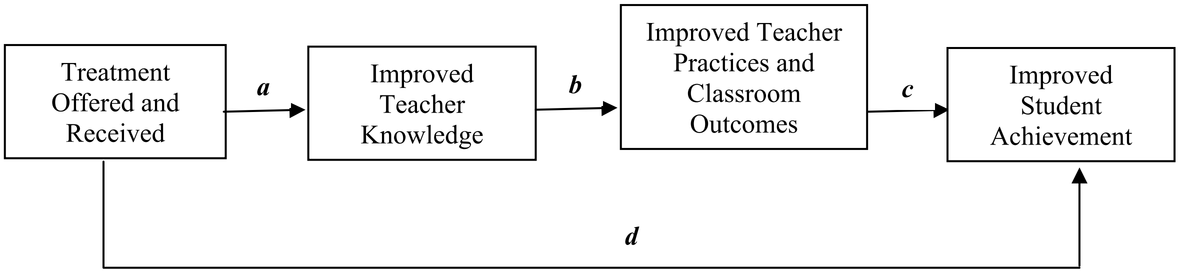
The rest of this report is in five chapters. Chapter 2 defines a "mediator" for the paper, and Chapter 3 discusses the theoretical framework for the analysis. Chapter 4 develops formulas for calculating statistical power using the OLS and IV regression frameworks, Chapter 5 presents the statistical power simulation results, and Chapter 6 presents a summary and conclusions.

Chapter 2: Definition of a Mediator

For the RCTs considered in this paper, a given classroom- or teacher-level variable is considered to be a mediator if it can partly account for the relationship between the offer of treatment services and student test scores (Baron and Kenny 1986). A mediator is an intermediate outcome that is measured *after* random assignment and that can be affected by the treatment.

To clarify, consider a typical conceptual model diagrammed in Figure 2.1 for an RCT of a teacher professional development intervention. In this path model, the causal chain is that the offer and receipt of intervention services first improves teacher knowledge (path ***a***), thereby improving teacher practices (path ***b***), and ultimately student test scores (path ***c***). In this model, teacher knowledge and practice measures are *mediating* outcomes that are measured for both the treatment and control groups. In some evaluations, the logic model may also have a *direct* link between treatment receipt and student test scores that is not via the teacher (path ***d***).

Figure 2.1: Typical Conceptual Model for an Education RCT


Treatment Offered and Received
a
Improved Teacher Knowledge
b
Improved Teacher Practices and Classroom Outcomes
c
Improved Student Achievement
d

The theoretical framework presented below develops statistical power formulas for estimating a generic mediator-achievement relationship. However, the primary focus of the empirical analysis is on teacher (classroom) practice mediators and the extent to which they mediate intervention effects on test scores. Stated differently, using Figure 2.1, the focus is on path ***ab***, the direct effect of offering the treatment on teacher practices, and path ***c***, the direct effect of teacher practices on student achievement, which is hereafter referred to as the "*mediator effect*." Teacher practice mediators are of particular importance for education RCTs, because they are expensive to collect and are typically considered to be key intermediate outcomes in the causal chain for improving student achievement. Thus, for simplicity, the teacher knowledge chain is ignored for the empirical analysis (or is assumed to be subsumed in the teacher practice chain). In addition, the empirical analysis does not consider mediators measuring the quality or amount of intervention services *received* by treatment teachers.

Chapter 3: Theoretical Framework

This chapter discusses the mathematical framework for the statistical power analysis, including the assumptions and basic regression models that are used for the analysis, and the general approach for calculating statistical power.

Assumptions

It is assumed that a multi-level RCT is conducted in *n* schools (indexed by *i*), with *c* classrooms per school (indexed by *j*) and *m* students per classroom (indexed by *k*). A balanced design is assumed, because it simplifies the variance formulas and cluster sample sizes are often similar for RCTs in the education area. However, the formulas presented below apply approximately for unbalanced designs if *c* and *m* are replaced by the average cluster sizes $\bar{c}$ and $\bar{m}$, respectively (Kish 1965).

It is assumed that schools are randomly assigned to a single treatment or control condition—the most common design used in education RCTs—where *p* is the sampling rate to the treatment group $(0 < p < 1)$. Thus, the sample contains np treatment and $n(1-p)$ control schools.

The study is assumed to take place during one school year, where an achievement test is administered to students in the fall and spring of the school year and continuous mediators (teacher practice measures) are collected in the spring of the school year. It is assumed that data are available for *both* treatment and control group students and teachers.

Student test scores are the focus of the analysis, because they are typically the key outcome for RCTs funded by the U.S. Department of Education and foundations. Although the conceptual model in Figure 2.1 posits a link between teacher practices and student test scores that are measured in *levels*, the analysis uses simple regression models to link the two outcomes using student test score *gains*. The use of gain scores (or alternatively, the inclusion of pretest scores as model covariates) yields more precise estimates of mediator-achievement associations than if the pretest scores were excluded from the analysis, and can adjust for differences between the abilities of students assigned to different classrooms that could bias these estimated relationships.

The regression analysis focuses on a *single* teacher practice mediator for several reasons. First, this is a reasonable starting point for a mediator analysis, where the relationship between test score gains and various mediators are looked at one at a time. Second, examining the statistical power of one mediator holding constant the effects of others would require additional ad hoc assumptions about correlations among the mediators included in the model. Third, for the IV approach, treatment status can be used as a valid instrument for only one mediator. Finally, the use of a single mediator simplifies the presentation and formulas, and is likely to yield empirical results that are suggestive of those from more complicated mediator analyses.

Estimation Models

Using the approach discussed in Baron and Kenny (1986), MacKinnon and Dwyer (1993), and Sobel (2008), the meditational hypotheses for the conceptual models considered in this paper can be tested as follows: (1) regress student test score gains on treatment status to estimate the average treatment effect (*ATE*) on student achievement, (2) regress the mediator on treatment status to estimate the *ATE* on the mediator, and (3) regress student test score gains on the mediator and treatment status to estimate the mediator effect. To establish mediation, the estimated *ATEs* and mediator effects must be nonzero and in the expected direction.

This section formalizes this framework for clustered RCTs in the education area. The considered mediator models are simple regression models that aim to associate *observable* student achievement and teacher practice outcomes. This paper does not consider more complex structural equation, path or latent variable models (see, for example, Kline 2005 and MacCallum et al. 1996); thus, the results presented here may not pertain to these approaches.

Impact Models

The *ATE* for student gain scores can be estimated using a random effects model or a hierarchical linear model (HLM) (Bryk and Raudenbush 1992):

$$(1) \quad y_{ijk} = \alpha_0 + \alpha_1 T_i + (u_i^y + \theta_{ij}^y + \varepsilon_{ijk}^y),$$

where y_{ijk} is the observed gain score for student *k* in classroom *j* and school *i*; T_i is 1 for treatments and 0 for controls; α_1 is the school-level *ATE* parameter; α_0 is the intercept; u_i^y are independently and identically distributed (*iid*) $N(0,\sigma_{uy}^2)$ school-level errors; θ_{ij}^y are *iid* $N(0,\sigma_{\theta y}^2)$ classroom-level errors; and ε_{ijk}^y are *iid* $N(0,\sigma_{\varepsilon y}^2)$ student-level errors. It is assumed that the error terms across levels are distributed independently of each other, and that the same error structure applies to both treatments and controls.

Importantly, the classroom-level error θ_{ij}^y reflects classroom-level variation in student test score gains, including both persistent and transitory effects (such as a "barking dog" effect that influences all students in the classroom at the time the test is administered). The literature provides separate estimates for these two effects using longitudinal student and teacher data and estimating models similar to (1) (see, for example, Goldhaber 2002; Hanushek et al. 2005; Jacob and Lefgren 2005; McCaffrey et al. 2004; Nye et al. 2004; and Rothstein 2009). Estimates of the persistent classroom effects may capture teacher-, student-, and school-related factors that influence student achievement. Thus, these estimated effects are hereafter referred to as estimated "*classroom effects*" rather than "teacher effects." As discussed below, published estimates on the extent to which the classroom-level variation in student test score gains explains the total variation in student gain scores plays a critical role for this paper. Note also from (1) that although the intervention might improve test scores, estimates of classroom effects are conditional on (net of) these impacts.

The *ATE* for the mediator can be estimated using the following model:

$$(2) \quad M_{ij} = \beta_0 + \beta_1 T_i + (u_i^M + \theta_{ij}^M),$$

where M_{ij} is the observed continuous mediator for teacher *j* in school *i*; β_1 is the *ATE* parameter for the mediator; and u_i^M and θ_{ij}^M are *iid* $N(0,\sigma_{uM}^2)$ and *iid* $N(0,\sigma_{\theta M}^2)$ random errors, respectively. It is assumed that $Cov(\theta_{ij}^M, u_i^M) = 0$. The errors across (1) and (2) could be correlated.

Mediator Models

The basic mediator model used for this paper is as follows:

$$(3)\quad y_{ijk} = \gamma_0 + \gamma_1 M_{ij} + \gamma_2 T_i + (u_i + \theta_{ij} + \varepsilon_{ijk}),$$

where M_{ij} is linked, by classroom, to each student; γ_0 is the intercept; γ_1 is the direct effect of the mediator on student gain scores (the mediator effect); γ_2 is the treatment effect on student gain scores due to school-related factors other than M_{ij}; and u_i, θ_{ij}, and ε_{ijk} are, respectively, *iid* $N(0,\sigma_u^2)$, *iid* $N(0,\sigma_\theta^2)$, and *iid* $N(0,\sigma_\varepsilon^2)$ random errors that are distributed independently of each other. The same error structure is assumed to apply to treatments and controls. Unlike (1), the errors in (3) are conditional on M_{ij}, and the inclusion of M_{ij} will typically reduce the variances of the classroom- and school-level errors, but not the variances of the student-level errors.

Note from (3) that:

$$(3a)\quad E(y_{ijk} \mid T_i = 1) - E(y_{ijk} \mid T_i = 0) = \gamma_1[E(M_{ij} \mid T_i = 1) - E(M_{ij} \mid T_i = 0)] + \gamma_2,$$

or, equivalently, that:

$$(3b)\quad \alpha_1 = \gamma_1 \beta_1 + \gamma_2,$$

where α_1 and β_1 are the *ATE* parameters in (1) and (2), respectively. Thus, the total effect of the intervention on y_{ijk} can be expressed as the sum of the indirect effect of the intervention on y_{ijk} via the mediator (that is, $\gamma_1\beta_1$ or path ***(ab)c*** in Figure 2.1) and the direct effect of the intervention on y_{ijk} due to other factors (that is, γ_2 or path ***d*** in Figure 2.1).

Note that because of random assignment, estimates of α_1 and β_1 have a causal interpretation. However, because M_{ij} values are self-selected, estimates of γ_1 and γ_2 may not have a causal interpretation except under certain conditions (see Sobel (2008) and below). Thus, the mediator effects, γ_1, are often referred to in this paper as "associations."

Using (3b), the estimated *ATE*s on M_{ij} and y_{ijk} can be linked by calculating the ratio $\hat{L} = \hat{\gamma}_1\hat{\beta}_1 / \hat{\alpha}_1$ (or $\hat{L}^{/} = (1-\hat{L}) = \hat{\gamma}_2 / \hat{\alpha}_1$), where $\hat{\gamma}_1$, $\hat{\beta}_1$, and $\hat{\alpha}_1$ are estimators for γ_1, β_1, and α_1, respectively. $\hat{L}$ is the proportion of the student-level impact that can be explained by the teacher-level impact, as posited by the study's conceptual model. Note that $\hat{L}$ is defined only if $\hat{\alpha}_1 \neq 0$, and will be nonzero only if $\hat{\beta}_1 \neq 0$ and $\hat{\gamma}_1 \neq 0$. An alternative approach is to consider only the numerator of $\hat{L}$, which is often referred to as the mediated effect (MacKinnon and Dwyer 1993).

The variance of $\hat{L}$ can be approximated using a standard Taylor series expansion of $\hat{L}$ around the true L and applying the delta method (Greene 2000). Focusing on first-order terms only (that is, ignoring covariance terms), this approach yields the following variance estimator:

$$(3c)\quad V\hat{a}r(\hat{L}) = \frac{\hat{\beta}_1^2\hat{\gamma}_1^2}{\hat{\alpha}_1^4}V\hat{a}r(\hat{\alpha}_1) + \frac{\hat{\gamma}_1^2}{\hat{\alpha}_1^2}V\hat{a}r(\hat{\beta}_1) + \frac{\hat{\beta}_1^2}{\hat{\alpha}_1^2}V\hat{a}r(\hat{\gamma}_1),$$

which can be used for significance testing.[1]

Although $\hat{L}$ can be used to gauge the merits of the conceptual model, for several reasons, this paper focuses more narrowly on examining statistical power for the mediator effect $\hat{\gamma}_1$. First, while $\hat{L}$ is a useful summary statistic for aggregating pieces of the conceptual model, it is desirable from a design perspective to have sufficient power for analyzing the strength of each piece of the chain separately. Stated differently, $\hat{L}$ is likely to be most informative if *each* of its components (that is, $\hat{\alpha}_1$, $\hat{\beta}_1$, and $\hat{\gamma}_1$) is estimated precisely. For example, it would be difficult to interpret a finding where $\hat{L}$ is statistically significant whereas some of its components are not due to low statistical power (which is theoretically possible). Second, most large-scale RCTs are designed to yield precise values of $\hat{\alpha}_1$ and $\hat{\beta}_1$, but rarely address statistical power for $\hat{\gamma}_1$. Thus, the goal of this paper is to identify appropriate sample sizes for obtaining precise estimates of $\hat{\gamma}_1$. Finally, from an empirical standpoint, it would difficult to conduct a "typical" statistical power analysis for $\hat{L}$, because its variance is a function of the unknown parameters α_1, β_1, and $\hat{\gamma}_1$, and there are no clear precision standards for L in education research. As discussed below, these problems can be overcome for a power analysis of $\hat{\gamma}_1$.

Note that there are two important features of (3). First, the model assumes the same mediator effect for treatments and controls (that is, the intervention is assumed to have a negligible effect on the mediator-achievement association). Second, school effects are treated as random in the model error term, so γ_1 is a weighted average of mediator effects *within* and *between* schools. The within-school component can be viewed as the parameter estimate of the mediator effect from a model where school-level means are subtracted from the data, whereas the between-school component can be viewed as the parameter estimate of the mediator effect from a model where the data are averaged to the school level.

This paper also considers the following variant of (3) that can be used to separately estimate the between- and within-school mediator effects:

$$(4)\quad y_{ijk} = \gamma_0 + \gamma_{1B}\bar{M}_i + \gamma_{1W}(M_{ij} - \bar{M}_i) + \gamma_2 T_i + (u_{1i} + \theta_{1ij} + \varepsilon_{1ijk}),$$

where $\bar{M}_i = \sum_{j=1}^{c} M_{ij} / c$ is the mean of M_{ij} in school *i*; γ_{1B} is the between-school mediator effect; γ_{1W} is the within-school mediator effect; and u_{1i}, θ_{1ij}, and ε_{1ijk} are *iid* normal random errors. In practice, estimates of γ_{1W} may be more defensible than estimates of γ_{1B}, because the between-school estimates may be more likely to suffer from omitted variable biases due to differences across schools in their environments, administrators, and student populations. For example, a positive estimate for γ_{1B} may not truly signify that schools with higher average test scores have better teachers if there are other school-

[1] Sobel (1982) presents a similar variance formula for the mediated effect $\hat{\gamma}_1\hat{\beta}_1$.

related factors—omitted from (4)—that partly account for these higher test scores. Note that (4) reduces to (3) if $\gamma_{1B} = \gamma_{1W}$.

Framework for Calculating Statistical Power

In this paper, (3) and (4) are used to examine statistical power for testing the null hypothesis $H_0 : \gamma_1 = 0$ versus the alternative hypothesis $H_1 : \gamma_1 \neq 0$. An F test is used for hypothesis testing using the statistic $t^2 = \hat{\gamma}_1^2 / V\hat{a}r(\hat{\gamma}_1)$, where $\hat{\gamma}_1$ is an estimator for γ_1. The test is to reject H_0 at significance level α if $t^2 \geq F_{1-\alpha}(1, n-1)$, which is the $(1-\alpha)th$ percentile of the *F* distribution with 1 degree of freedom for the numerator and $(n-1)$ degrees of freedom for the denominator.

The statistical power of this test—the probability of rejecting H_0 given that H_1 is true— can be computed using the non-central *F* distribution:

$$(5) \quad \Pr\{F(1, n-1, \delta) \geq F_{1-\alpha}(1, n-1)\},$$

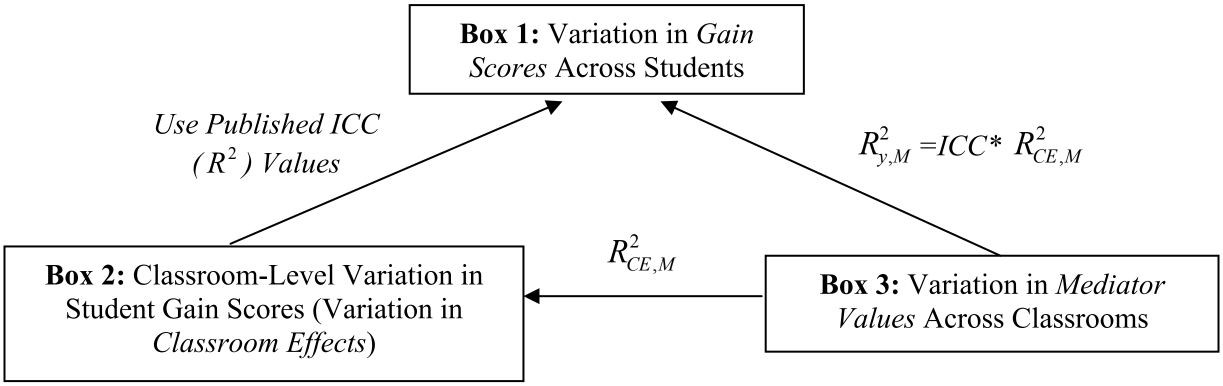
where the non-centrality parameter δ is defined as follows:

$$(6) \quad \delta = E(\hat{\gamma}_1)^2 / Var(\hat{\gamma}_1),$$

where $E(\hat{\gamma}_1) = \gamma_1$ for an unbiased estimator. The parameter δ depends on the size of $E(\hat{\gamma}_1)$ and the variance of $\hat{\gamma}_1$, which is a function of study sample sizes and design effects due to clustering. Statistical power is determined by δ and increases as δ increases. Thus, the focus of the theoretical analysis is to develop formulas for δ for various OLS and IV estimators. In the empirical analysis, these formulas are inserted into (5) to identify minimum values for γ_1 to ensure that an RCT will have a high probability (say, 80 percent) of finding a statistically significant mediator-achievement association.

To help interpret these γ_1 values, the formulas for δ are instead expressed in terms of *population* R^2 *values*, or the proportion of variance in student gain scores that must be explained by the variation in the mediator (see Figure 3.1 and Nye et al, 2004). This metric is useful because plausible R^2 values ($R^2_{y,M}$ values in Figure 3.1) can be obtained using published intraclass correlation (*ICC*) estimates on the extent to which the variance in test score gains can be explained by the variance in estimated *classroom effects* as defined above (see the link between Boxes 1 and 2 in Figure 3.1). These *ICC*s are likely to provide an upper bound on $R^2_{y,M}$ values, because, in practice, it is likely that the variation in the mediators will explain only part of the classroom-level variation in student gain scores (as determined by $R^2_{CE,M}$ values in Figure 3.1). These $R^2_{CE,M}$ values are likely to be small due to limitations on the dimensions of teacher practices that can be captured by the mediators and measurement error in the mediators. Plausible values for *ICC*, $R^2_{y,M}$, and $R^2_{CE,M}$ are discussed below for the empirical analysis.

Figure 3.1: The Use of Regression R^2 Values for the Power Analysis


Box 1: Variation in *Gain Scores* Across Students
Use Published ICC (R^2) Values
$R^2_{y,M} = ICC^* R^2_{CE,M}$
$R^2_{CE,M}$
Box 2: Classroom-Level Variation in Student Gain Scores (Variation in *Classroom Effects*)
Box 3: Variation in *Mediator Values* Across Classrooms

Chapter 4: Statistical Power Formulas

This chapter presents formulas for the non-centrality parameter in (6) using OLS and IV methods for estimating the mediator associations in (3) and (4). Asymptotic formulas are presented due to the considerable complexity of calculating finite sample moments for IV estimators. OLS estimators are used rather than GLS estimators (which are more efficient), because the IV approach and corrections for mediator measurement error are much more complex using the GLS approach.

OLS Approach Using the Control Group Only

To fix ideas, consider the estimation of γ_1 in (3) using OLS methods and the control group only, so that $T_i = 0$ for all observations (this scenario is also pertinent for studies that collect mediator and achievement data but that are not impact evaluations). Under this scenario, OLS will produce consistent estimators if $Cov(M_{ij}, u_i) = Cov(M_{ij}, \theta_{ij}) = Cov(M_{ij}, \varepsilon_{ijk}) = 0$, that is, if the error terms in (2) and (3) are uncorrelated. This will occur under three conditions: (1) the model error terms do not include omitted variables that are correlated with the mediator; (2) the mediator cannot be determined simultaneously with student gain scores (that is, it cannot be the case that teachers who teach easier-to-serve, more motivated students at the outset have higher values for the mediator); and (3) there is no measurement error in M_{ij}.[2] These orthogonality assumptions are probably unrealistic, but the OLS approach is a reasonable starting point for a mediator analysis

The OLS estimator for γ_1 in (3) is:

$$(7) \quad \hat{\gamma}_{1,OLSa} = \frac{\sum_{i=1}^{n(1-p)} \sum_{j=1}^{c} \sum_{k=1}^{m} (y_{ijk} - \bar{y}_{...})(M_{ij} - \bar{M}_{...})}{m \sum_{i=1}^{n(1-p)} \sum_{j=1}^{c} (M_{ij} - \bar{M}_{...})^2},$$

where $\bar{y}_{...}$ and $\bar{M}_{...}$ are grand control group means for y_{ijk} and M_{ij}, respectively. Standard OLS methods can be used to show that under the orthogonality assumptions discussed above, $\hat{\gamma}_{1,OLSa} \xrightarrow{p} \gamma_1$ as *n* approaches infinity (for fixed *c* and *m*), where $\xrightarrow{p}$ denotes convergence in probability.

To derive the asymptotic variance of $\hat{\gamma}_{1,OLS}$, define $\mathbf{X_i} = [\mathbf{1}\ \mathbf{M_i}]$ as the $(cm)x2$ matrix of model covariates for students in school *i*, where $\mathbf{1}$ is a column vector of 1s and $\mathbf{M_i}$ is a column vector containing the M_{ij} s. In addition, define $\mathbf{\Omega_i}$ as the $(cm)x(cm)$ variance-covariance matrix for students

[2] Holland (1988) and Sobel (2008) discuss these conditions in terms of potential mediator values $M(T_i)$, and potential student outcomes, $Y(T_i, m)$, where *m* denotes possible mediator values. The key condition is that $M(T_i)$ is *ignorable* with respect to (independent of) $Y(T_i, m)$ for all *m* and for $T_i = 0,1$.

in school i, whose diagonal elements are $\sigma^2 = (\sigma_u^2 + \sigma_\theta^2 + \sigma_\varepsilon^2)$, and whose off-diagonal elements are $(\sigma_u^2 + \sigma_\theta^2)$ for students in the same classroom and σ_u^2 for students in different classrooms. The variance of $\hat{\gamma}_{1,OLSa}$ can then be expressed as follows:

$$(8) \quad Var(\hat{\gamma}_{1,OLSa}) = [(\sum_{i=1}^{n(1-p)} \mathbf{X_i'X_i})^{-1}(\sum_{i=1}^{n(1-p)} \mathbf{X_i'\Omega_iX_i})(\sum_{i=1}^{n(1-p)} \mathbf{X_i'X_i})^{-1}]_{2,2}.$$

After applying some algebra to (8) and taking probability limits, the asymptotic variance of $\hat{\gamma}_{1,OLSa}$ becomes:

$$(9) \quad AsyVar(\hat{\gamma}_{1,OLSa}) = \frac{\sigma^2}{n(1-p)cm\sigma_M^2}[1 + \rho_1(m-1) + \rho_2(cm\psi - 1)],$$

where $\rho_1 = \sigma_\theta^2 / \sigma^2$ is the classroom-level population *ICC* from (3); $\rho_2 = \sigma_u^2 / \sigma^2$ is the school-level population *ICC* from (3); $\psi = \sigma_{\bar{M}_B}^2 / \sigma_M^2$; and $\sigma_M^2 = \sigma_{uM}^2 + \sigma_{\theta M}^2$ and $\sigma_{\bar{M}_B}^2 = \sigma_{uM}^2 + (\sigma_{\theta M}^2 / c)$ are population variances of M_{ij} and $\bar{M}_i$, respectively. Note that the *ICC* for the mediator is $\rho_M = \sigma_{uM}^2 / \sigma_M^2$, and thus, $\psi = \rho_M + [(1-\rho_M)/c]$. The OLS estimator $\hat{\gamma}_{1,OLSa}$ is asympotically normally distributed (see, for example, Rao 1973).

The variance formula in (9) is the product of the variance of the simple OLS estimator and the *design effect* (in brackets) due to the clustering of students within classrooms and schools. The design effect will be small if ρ_1 and ρ_2 are small, which will occur if student gain scores conditional on the mediator vary little across classrooms and schools. Design effects also become smaller as ψ becomes smaller.

The variance in (9) can be expressed in terms of the population squared correlation between student gain scores and the mediator, $R_{y,M}^2$, by noting that: (1) $R_{y,M}^2 = \sigma_{y,M}^2 /(\sigma_y^2\sigma_M^2) = \gamma_1^2\sigma_M^2 / \sigma_y^2$, where $\sigma_{y,M}$ is the population covariance between y_{ijk} and M_{ij}; and (2) $\sigma^2 = \sigma_y^2(1 - R_{y,M}^2)$, where $\sigma_y^2 = \sigma_{uy}^2 + \sigma_{\theta y}^2 + \sigma_{\varepsilon y}^2$. Using these relations, the asymptotic non-centrality parameter in (6) can be expressed as follows:

$$(10) \quad \delta_{OLSa} = \frac{\gamma_1^2}{AsyVar(\hat{\gamma}_{1,OLSa})} = \frac{n(1-p)cmR_{y,M}^2}{(1-R_{y,M}^2)deff},$$

where $deff = [1 + \rho_1(m-1) + \rho_2(cm\psi - 1)]$. This is the clustered version of the non-centrality parameter for regression coefficients under non-clustered designs (see, for example, Cohen 1977, 1988). It is intuitive that the design effect will typically reduce the value of the non-centrality parameter, which leads to reductions in statistical power.

Finally, similar methods can be used to show that the asymptotic non-centrality parameters for the (orthogonal) between- and within-school mediator-test score associations in (4) are as follows:

$$(11) \quad \delta_{B,OLSa} = \frac{n(1-p)cmR^2_{y,\bar{M}_B}}{(1-R^2_{y,\bar{M}_B} - R^2_{y,M_W})deff_B}, \text{ and}$$

$$(12) \quad \delta_{W,OLSa} = \frac{n(1-p)cmR^2_{y,M_W}}{(1-R^2_{y,\bar{M}_B} - R^2_{y,M_W})deff_W},$$

where $R^2_{y,\bar{M}_B}$ is the population squared correlation between y_{ijk} and $\bar{M}_i$, R^2_{y,M_W} is the population squared correlation between y_{ijk} and $(M_{ij} - \bar{M}_i)$, $deff_B = [1 + \rho_1(m-1) + \rho_2(cm-1)]$, and $deff_W = [1 + \rho_1(m-1)]$.

OLS Approach Using the Control Group With Measurement Error

Thus far, it has been assumed that the mediator is measured without error. However, as discussed in Raudenbush et al. (2008), mediator measurement error can be large, especially for classroom observation measures. This is because classroom observation data are typically collected by a small number of raters during a few short time intervals, and there can be considerable variation in measurement across raters, in the quality of teacher practices during the day, and from interactions between these factors.

Measurement error is incorporated into the analysis using a standard measurement error model (see, for example, Fuller 1987):

$$(13) \quad M_{ij}^{Obs} = M_{ij} + \xi_{ij},$$

where M_{ij}^{Obs} is the *observed* mediator and ξ_{ij} is an *iid* $N(0,\sigma^2_\xi)$ random measurement error term that is uncorrelated with M_{ij} and the other error terms in (1)-(4). The error term ξ_{ij} could include random effects such as rater and segment effects, and thus, σ^2_ξ includes these sources of variation (Raudenbush et al. 2008). Using (2) and (13), the *reliability* of the mediator is defined as follows:

$$(13a) \quad \lambda_{rel} = \sigma^2_{\theta M} / (\sigma^2_{\theta M} + \sigma^2_\xi).$$

Consider the estimation of (3) in the presence of measurement error. In this case, the "true" model is (3), but the estimation model is:

$$(14) \quad y_{ijk} = \gamma_0 + \gamma_1 M_{ij}^{Obs} + (u_i + \theta_{ij} + \varepsilon_{ijk} - \gamma_1 \xi_{ij}).$$

In this model, M_{ij}^{Obs} is correlated with the error term because $E(M_{ij}^{Obs}\xi_{ij}) = \sigma^2_\xi$. The resulting OLS estimator, $\hat{\gamma}_{1,OLSa,ME}$, suffers from attenuation bias because $\hat{\gamma}_{1,OLSa,ME} \xrightarrow{p} \lambda\gamma_1$, where $0 \le \lambda = \sigma^2_M / (\sigma^2_M + \sigma^2_\xi) \le 1$. Note that λ is greater than the reliability of the mediator, λ_{rel}, because $\sigma^2_M = \sigma^2_{\theta M} + \sigma^2_{uM} > \sigma^2_{\theta M}$. With measurement error, a consistent OLS estimator is $(\hat{\gamma}_{1,OLSa,ME} / \hat{\lambda})$, where $\hat{\lambda}$ is an estimate of λ (which is often difficult to obtain in practice).

Using results in Fuller (1987), the asymptotic variance of $\hat{\gamma}_{1,OLSa,ME}$ is as follows:

$$(15)\quad AsyVar(\hat{\gamma}_{1,OLSa,ME}) = \frac{\sigma_y^2(1-R^2_{y,M^{Obs}})deff_1}{ncm\sigma^2_{M^{Obs}}} = \frac{\lambda\sigma_y^2(1-\lambda R^2_{y,M})deff_1}{ncm\sigma^2_M},$$

where $\sigma^2_{M^{Obs}} = \sigma^2_M + \sigma^2_\xi$ is the population variance of M^{Obs}_{ij}, $R^2_{y,M^{Obs}}$ is the population squared correlation between y_{ijk} and M^{Obs}_{ij}, $deff_1 = [1+\rho_1(m-1)+\rho_2(cm\psi^{Obs}-1)]$ is the design effect, $\psi^{Obs} = (\sigma^2_{\bar{M}^{Obs}_B} / \sigma^2_{M^{Obs}}) \le \psi$, and $\sigma^2_{\bar{M}^{Obs}_B} = \sigma^2_{\bar{M}_B} + (\sigma^2_\xi / c)$. The second equality in (15) holds because $\sigma^2_{M^{Obs}} = \sigma^2_M / \lambda$ and $R^2_{y,M^{Obs}} = \lambda R^2_{y,M}$.

Using (15), the non-centrality parameter with measurement error becomes:

$$(16)\quad \delta_{OLSa} = \frac{(\lambda\gamma_1)^2}{AsyVar(\hat{\gamma}_{1,OLSa,ME})} = \frac{n(1-p)cm\lambda R^2_{y,M}}{(1-\lambda R^2_{y,M})deff_1},$$

where the second equality holds because $R^2_{y,M} = \gamma_1^2\sigma^2_M / \sigma^2_y$. Thus, measurement error reduces the non-centrality parameter (and hence, statistical power) by lowering the model R^2 values. Intuitively, statistical power is lower in the presence of measurement error, because it becomes more difficult for the data to isolate the signal from the noise in the observed mediator.

Measurement error also biases the estimated between- and within-school associations in (4), but by different factors. Specifically, $\hat{\gamma}_{1B,OLSa,ME} \xrightarrow{p} \lambda_B\gamma_{1B}$ where $\lambda_B = \sigma^2_{\bar{M}_B} / [\sigma^2_{\bar{M}_B} + (\sigma^2_\xi / c)]$, and $\hat{\gamma}_{1W,OLSa,ME} \xrightarrow{p} \lambda_{rel}\gamma_{1W}$ where λ_{rel} is the reliability of the mediator from above. Thus, with measurement error, the non-centrality parameters in (11) and (12) can be updated by replacing $R^2_{y,\bar{M}_B}$ with $\lambda_B R^2_{y,\bar{M}_B}$ and R^2_{y,M_W} with $\lambda_{rel} R^2_{y,M_W}$.

OLS Approach Using the Treatment and Control Groups

Under the orthogonality assumptions discussed above (conditional on T_i), OLS also produces consistent estimates of the mediator associations in (3) and (4) when the estimation sample includes both treatments and controls. As shown in (3b), these models decompose the total *ATE* on student gain scores into a part due to the mediator and another part due to residual school-related factors (represented by γ_2). Note that $\bar{M}_i$ and T_i may be correlated (which complicates the analysis), but not $(M_{ij} - \bar{M}_i)$ and T_i.

The estimation of (3) and (4) using the full sample can be performed using similar methods to those using the control group only. For simplicity, this section uses the same notation as above, but parameters such as σ^2_M, $\sigma^2_{\bar{M}_B}$, ψ, σ^2_y, and λ are now *unconditional* on treatment status. For example, using (2), σ^2_M is now $[\beta_1^2 p(1-p)+\sigma^2_{\theta M}+\sigma^2_{uM}]$ rather than $[\sigma^2_{\theta M}+\sigma^2_{uM}]$.

Let $\mathbf{X_i}=[\mathbf{1}\ \mathbf{M_i}\ \mathbf{Q_i}]$ be the new covariate matrix, where $\mathbf{Q_i}$ is a $(cm)x1$ column vector containing the T_is, and let $\mathbf{y_i}$ be the vector of student gain scores in school *i*. The OLS estimator is then $\hat{\gamma}_{1,OLSb}=[(\sum_i \mathbf{X}'_i\mathbf{X_i})^{-1}\sum_i \mathbf{X'_i y_i}]_{2,2}$, which is consistent and asymptotically normal.

Using (8) and taking probability limits, an approximation to the asymptotic variance of $\hat{\gamma}_{1,OLSb}$ is as follows:

$$(17)\quad AsyVar(\hat{\gamma}_{1,OLSb})\approx\frac{\sigma^2 deff}{ncm\sigma^2_M(1-R^2_{M,T})},$$

where $R^2_{M,T}=\psi R^2_{\bar{M}_B,T}$ is the squared population correlation between M_{ij} and T_i, and other terms are defined as above.[3]

Similar to the case above, measurement error will result in downwardly biased OLS estimates, because $\hat{\gamma}_{1,OLSb,ME}\xrightarrow{p}\lambda\gamma_1$, where $\lambda=\sigma^2_M(1-R^2_{M,T})/\sigma^2_{M^{Obs}}(1-R^2_{M^{Obs},T})$. Thus, as shown in Appendix A, the resulting asymptotic non-centrality parameter for $\hat{\gamma}_{1,OLSb,ME}$ is:

$$(18)\quad \delta_{OLSb}\approx\frac{ncm\lambda R^2_{y,M|T}}{(1-\lambda R^2_{y,M|T})deff_1},$$

where $R^2_{y,M|T}$ is the squared *partial* correlation between y_{ijk} and M_{ij}, controlling for T_i.

Finally, similar methods reveal that the corresponding asymptotic non-centrality parameters for the between- and within-school mediator effects in (4) are as follows:

$$(19)\quad \delta_{B,OLSb}=\frac{ncm\lambda_B R^2_{y,\bar{M}_B|T}}{(1-\lambda_B R^2_{y,\bar{M}_B|T}-\lambda_{rel}R^2_{y,M_W})deff_B},\text{ and}$$

$$(20)\quad \delta_{W,OLSb}=\frac{ncm\lambda_{rel} R^2_{y,M_W}}{(1-\lambda_B R^2_{y,\bar{M}_B|T}-\lambda_{rel}R^2_{y,M_W})deff_W},$$

[3] The actual asymptotic variance is $\frac{\sigma^2 p(1-p)[p(1-p)\sigma^2_M deff-\sigma^2_{M,T}deff_B]}{ncm[p(1-p)\sigma^2_M-\sigma^2_{M,T}]}$, where $\sigma_{M,T}$ is the population covariance between M_{ij} and T_i. This variance reduces to (17) assuming that $deff_B=deff$ and using the relation $R^2_{M,T}=\sigma^2_{M,T}/[\sigma^2_M p(1-p)]$.

where $R^2_{y,\bar{M}_B|T}$ is the squared partial correlation between y_{ijk} and $\bar{M}_i$, controlling for T_i.

Instrumental Variables Approach Using the Treatment and Control Groups

As discussed, the OLS approach considered above will yield biased estimates in the presence of simultaneity and omitted variable biases. Mediator measurement error could also lead to biased OLS estimators.

Under certain assumptions, an IV approach using the full sample—that exploits the experimental design—can be used to adjust for these potential biases and produce consistent estimates (see, for example, Bloom et al. 2009; Holland 1988; Sobel 2008; and Wooldridge 2002). In our context, the IV approach involves estimating (4), where γ_{1W} and γ_2 are set to zero, and where T_i is used as an instrument for $\bar{M}_i$. The estimation model under this approach is:

$$(21) \quad y_{ijk} = \gamma_0 + \gamma_{1B}\bar{M}_i + (u_{2i} + \theta_{2ij} + \varepsilon_{2ijk}),$$

where u_{2i}, θ_{2ij}, and ε_{2ijk} are normally distributed error terms (that exclude measurement error) with total variance σ^2_{IV}.[4]

There are two key conditions that are required for the consistency of the IV estimator. The first is that T_i must be uncorrelated with u_{2i}, θ_{2ij}, and ε_{2ijk} in (21) (see Angrist et al., 1996). This exclusion restriction implies that any effect of T_i on student gain scores must occur *only* through an effect of T_i on $\bar{M}_i$. This rules out alternative school-related mediating pathways through which the intervention can influence student learning (that is, path ***d*** in Figure 2.1 cannot exist). The plausibility of this assumption will depend on the particular intervention. For instance, it may hold for a teacher mentoring program where student learning gains are likely to be fully mediated through the teacher, but it may not hold if the intervention involves new computers in the classroom so that the treatment can affect student achievement through means other than improvements in teacher practices.

The second key condition required for the consistency of the IV estimator is that there must be a nonzero covariance between T_i and the model covariates. This implies that with school-based random assignment, the IV approach can only be used to identify mediator associations *between* schools (that is, at the school level), but not within schools. Furthermore, because $Cov(T_i, \bar{M}_i) = p(1-p)\beta_1$, this condition requires that the treatment effect on $\bar{M}_i$ (that is, β_1) must be nonzero. In the empirical work, it is assumed that β_1 is large enough so that T_i is a "strong" instrument (see Murray 2006 and Stock et al. 2002), although the weak instrument issue is a finite sample problem, and thus, does not affect the asymptotic formulas presented below.

[4] The variances of these error terms are assumed to be homoscedastic conditional on the instrument, T_i, rather than on the mediator.

To help understand the IV estimator, it is important to first consider the IV parameter that can be identified. Note from (21) that $\sigma_{y,T} = \gamma_{1B}\sigma_{\bar{M}_B,T}$, where $\sigma_{y,T}$ is the population covariance between y_{ijk} and T_i, and $\sigma_{\bar{M}_B,T}$ is the population covariance between $\bar{M}_i$ and T_i. Hence, this relation implies that the identifiable IV parameter is $\gamma_{1B} = \sigma_{y,T} / \sigma_{\bar{M}_B,T}$.

A consistent IV estimator, $\hat{\gamma}_{1B\_IV}$, is as follows:

$$(22) \quad \hat{\gamma}_{1B,IV} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{c}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{...})(T_i - p)}{mc\sum_{i=1}^{n}(\bar{M}_i - \bar{M}_{...})(T_i - p)} \xrightarrow{p} \frac{\sigma_{y,T}}{\sigma_{\bar{M}_B,T}} = \frac{\alpha_1}{\beta_1} = \gamma_{1B},$$

where all terms are defined as above. This estimator has a clear interpretation: it is the ratio of the school-level *ATE* on student test score gains and the school-level *ATE* on the mediator. Intuitively, $\hat{\gamma}_{1B,IV}$ represents the extent to which student tests scores improve due to an *exogenous* treatment-induced change in the mediator. IV estimators are known to be asymptotically normal under weak regularity conditions (see Wooldridge 2002).

The exposition above for the IV estimator assumes that treatment and mediator effects are *constant* across observations, so that the IV approach yields estimates of causal effects that pertain to the full study population (Holland 1988; Imbens and Angrist 1994; Sobel 2008). If schools respond differently to the treatment, however, the IV estimator can only recover a weighted average of local average treatment effects (LATEs) for the subpopulations affected by the treatment, where the weights are largest for those groups that respond most to the treatment.

The variance of the IV estimator, $\hat{\gamma}_{1B,IV}$, can be expressed as follows:

$$(23) \quad Var(\hat{\gamma}_{1B,IV}) = [(\sum_{i=1}^{n}\mathbf{Z_i'X_i})^{-1}(\sum_{i=1}^{n}\mathbf{Z_i'\Omega_i Z_i})(\sum_{i=1}^{n}\mathbf{Z_i'X_i})^{-1}]_{2,2},$$

where $\mathbf{Z_i} = [\mathbf{1}\ \mathbf{Q_i}]$ is a $(cm)x2$ matrix of instruments for the covariate matrix $\mathbf{X_i} = [\mathbf{1}\ \bar{\mathbf{M}}_\mathbf{i}]$. After some algebra, the probability limit of (23) becomes:

$$(24) \quad AsyVar(\hat{\gamma}_{1B,IV}) = \frac{\sigma^2_{IV}deff_B}{ncm\sigma^2_{\bar{M}_B}R^2_{\bar{M}_B,T}} = \frac{\sigma^2_y(1 - R^2_{y,\bar{M}_B,IV})deff_B}{ncm\sigma^2_{\bar{M}_B}R^2_{\bar{M}_B,T}},$$

where $R^2_{y,\bar{M}_B^{Obs},IV}$ is the population squared correlation in the IV model between y_{ijk} and $\bar{M}_i$.

The key difference between the IV and OLS variances for $\hat{\gamma}_{1B}$ is that the denominator in (24) contains $R^2_{\bar{M}_B,T}$, compared to $(1 - R^2_{\bar{M}_B,T})$ for the OLS estimator. $R^2_{\bar{M}_B,T}$ values are likely to be small, because treatment status is likely to explain only a small percentage of the total variance in the mediators (see below). Thus, a key finding is that the variance of the IV estimator is likely to be considerably *larger* than the variance of the comparable OLS estimator.

With measurement error, the IV estimator, $\hat{\gamma}_{1B,IV}$, remains consistent (unlike the OLS estimator). Furthermore, measurement error does not affect the denominator in (24), because $\sigma^2_{\bar{M}_B^{Obs}} = (\sigma^2_{\bar{M}_B} / \lambda_B)$ and $R^2_{\bar{M}_B^{Obs},T} = \lambda_B R^2_{\bar{M}_B,T}$. Thus, measurement error will only affect the variance of the mediator effect through $R^2_{y,\bar{M}_B^{Obs},IV}$, which can be expressed as follows:

$$(25) \quad R^2_{y,\bar{M}_B^{Obs},IV} = \frac{\sigma^2_{y,\bar{M}_B^{Obs}}}{\sigma^2_y \sigma^2_{\bar{M}_B^{Obs}}} = \frac{(\gamma_{1B}\sigma^2_{\bar{M}_B} + \sigma_{\bar{M}_B^{Obs},error})^2}{\sigma^2_y (\sigma^2_{\bar{M}_B} / \lambda_B)}.$$

In this expression, $\sigma_{\bar{M}_B,error}$ is the population covariance between $\bar{M}_i$ and the error terms in (21) (that exclude measurement error). The sign and magnitude of $\sigma_{\bar{M}_B^{Obs},error}$ will depend on specific study features, such as the study mediators and achievement tests, the nature of the intervention, and the study population. Because of this uncertainty, it is assumed for the empirical analysis that $\sigma_{\bar{M}_B^{Obs},error} = 0.$ In this simplifying case, $R^2_{y,\bar{M}_B^{Obs},IV} = \lambda_B R^2_{y,\bar{M}_B}$, and the non-centrality parameter for $\hat{\gamma}_{1B,IV,ME}$ is as follows:

$$(26) \quad \delta_{B,IV} = \frac{\gamma^2_{1B}}{AsyVar(\hat{\gamma}_{1B,IV,ME})} = \frac{ncmR^2_{y,\bar{M}_B} R^2_{\bar{M}_B,T}}{(1 - \lambda_B R^2_{y,\bar{M}_B}) deff_B}.$$

Although appealing at first glance, the IV approach has several limitations that could reduce its utility in school-based RCTs. First, the main effect, T_i, can be used as an instrument for only one mediator. The IV approach can be extended to the case of multiple mediators if there is variation in mediator impacts across exogenous subgroups, such as sites. In these cases, treatment-by-site interaction terms could be used as instruments for specific mediators (see, for example, Kling et al. 2007). However, to the extent that these instruments can be found, they may be weak instruments if the variation in mediator impacts across sites is limited (which may be the case in education RCTs). Weak instruments are a problem because they lead to IV estimators that are biased towards the OLS estimators (see Stock et al. 2002). Second, the variances of IV estimators are likely to be large, suggesting that mediator analyses using the IV approach will have low power (see below). Third, as discussed, because T_i is a school-level variable, the IV approach can only estimate mediator effects between schools, not within schools. Finally, the IV estimator provides causal effects for the full population only under certain conditions, such as constant treatment effects (or the slightly weaker conditions discussed in Sobel 2008).

Chapter 5: Empirical Analysis

This chapter uses the non-centrality parameter formulas in (18), (19), (20) and (26) to conduct a simulated "typical" statistical power analysis for γ_1 for school-based RCTs. The goal is to identify the number of study schools that are required to ensure that an RCT has sufficient power for detecting R^2 values for mediator effects that are likely to be found in practice. The focus is on RCTs for elementary school students in low performing schools, a common target population for experiments conducted in education research.

The first part of this chapter discusses the key issue of identifying benchmark R^2 values that can realistically be found in practice using the approach displayed in Figure 3.1 above. The second part discusses additional assumptions that are required for the statistical power calculations, and the final part presents the empirical results.

Identifying Plausible R^2 Values

To obtain benchmark R^2 values for the analysis, it is convenient to use estimates found in the literature on the proportion of the total variance in student gain scores that is due to classroom-level variation in gain scores—the ρ_1 and ρ_2 parameters from above (and the *ICC* parameters in Figure 3.1). As discussed, these *ICC*s are likely to provide an upper bound on the extent to which classroom-level mediators can explain the variation in student gain scores.

Chiang (2009) presents a host of *ICC* estimates from the literature and using new data sources. The estimates pertain to fall-spring test score gains on various math, reading, and language arts tests for elementary school students. Most studies were performed in low-income schools, but not all.

The *ICC*s in Chiang (2009) vary across studies, reflecting differences in study samples and achievement tests. The *ICC*s at the classroom level range from 0.02 to 0.15, and the *ICC*s at the school-level range from 0.05 to 0.20. Using mean values of $\rho_1 = 0.05$ and $\rho_2 = 0.10$, it appears that overall, about 15 percent of the variance in student gain scores can be explained by differences in classroom effects within and between schools.

A measured mediator can be expected to capture only particular dimensions of teacher practices, and thus, to explain only a fraction of the 15 percent variation in classroom effects within and between schools (this fraction is denoted by $R^2_{CE,M}$ in Figure 3.1). For example, Jacob and Lefgren (2005) found that principal assessments of teachers explained only about 10 percent of the variation in classroom effects on reading and math. Similarly, Aaronson et al. (2007) found that a host of teacher characteristics—including age, gender, race, educational background, tenure, and total experience—*together* only explained about 20 percent of the variation in classroom effects. Thus, it is likely that even a strong predictor of classroom effects could explain only a portion of this variation. Furthermore, mediator subscales, that can help determine which practices matter, may explain even less.

Based on this literature, the power calculations were conducted assuming that the mediator explains 10 percent of the 15 percent variation in classroom effects (that is, $R^2_{CE,M} = .10$ in Figure 3.1). This implies a benchmark R^2 value of 1.5 percent for the mediator effect γ_1 (which can be obtained using the relation $R^2_{y,M} = ICC * R^2_{CE,M} = .15 * .10$ in Figure 3.1). The calculations were also conducted using a more optimistic R^2 value of 3 percent ($R^2_{CE,M} = .20$), and a less optimistic R^2 value of .75 percent ($R^2_{CE,M} = .05$). Similarly, using values of $\rho_1 = 0.05$ and $\rho_2 = 0.10$, the power calculations assumed target R^2 values of 0.005, 0.01, and 0.0025 for the analysis of mediator effects within schools (γ_{1W}), and 0.01, 0.02, and 0.005 for the analysis of mediator effects between schools (γ_{1B}).

Finally, viewing these target R^2 values as *squared correlations* suggests also that they are nontrivial. For instance, the assumption that the mediator can explain 10 percent of the variance in estimated classroom effects implies a *correlation* of 0.32 between these two measures. Similarly, the assumption that the mediator can explain 20 percent of the variance in estimated classroom effects implies a correlation of 0.45, which is larger than those that are typically found in practice (Perez-Johnson et al. 2009).

Additional Assumptions for the Statistical Power Calculations

The statistical power calculations were conducted using the following "real-world" assumptions: (1) a two-tailed test, (2) a 5 percent significance level, (3) a balanced allocation of schools to the treatment and control groups ($p = 0.5$), (4) an average of 3 classrooms per school ($c = 3$), (5) an average of 23 students per classroom, (6) data on student test score gains are available for 80 percent of students in the sample (so that $m = 18.2$), and (7) data on mediating outcomes are available for all teachers.

The statistical power calculations also required real-world assumptions on values for several additional parameters that enter the non-centrality parameter formulas, as discussed next.

Reliability-Related Parameters (λ_{rel}, λ, and λ_B). The reliability of a teacher practice mediator, λ_{rel} as defined in (13a), will likely depend on the nature of the mediator and the study design. For example, reliability may differ for a mediator constructed using classroom observation data, principal ratings, or teacher survey data. Because of this uncertainty, the power calculations were conducted assuming reliability values of 0.2, 0.5, and 1.0. Although perfect reliability is never attainable, reliability values of 1 are used in the analysis as a best-case scenario.

The 0.2 and 0.5 values are in the range of plausible values for λ_{rel} reported in Raudenbush et al. (2008) based on an analysis of Classroom Assessment Scoring System (CLASS) data. Raudenbush et al. (2008) estimated the measurement error variances in (13) using the observed variation in instructional climate scores across raters and time segments. The 0.2 to 0.5 reliability values are lower than those usually reported for commonly-used classroom observation protocols. This is because the reliability values found in the literature are typically based on the internal consistency of item responses, and do not typically address the critical sources of measurement variation examined by Raudenbush et al. (2008).

Finally, for simplicity, the same parameter values are used for λ, λ_{rel}, and λ_B, even though these parameters may differ in practice.

The ratios ψ and ψ^{Obs}. These parameters represent the extent to which mean mediator values vary across schools, and enter the design effect formulas. As discussed, these parameters can be obtained from *ICC* estimates for the mediator. These *ICC*s, however, are not typically reported in study reports, and there is no literature that collates such *ICC* estimates from previous studies. Thus, to obtain plausible *ICC* values, classroom observation mediators were analyzed from two large school-based education RCTs: (1) the Evaluation of the Effectiveness of Selected Supplemental Reading Comprehension Interventions

(James-Burdumy et al. 2009), and (2) the Evaluation of Comprehensive Teacher Induction Programs (Glazerman et al. 2008). The Reading Comprehension study used the Expository Reading Comprehension (ERC) Classroom Observation Instrument, and the Teacher Induction study used the Vermont Classroom Observation Tool (Saginor and Hyjek 2005).

The *ICC* estimates for the mediators differ for the two studies. The *ICC* estimates for the Reading Comprehension study are 0.21 for the interactive teaching scale, 0.33 for the strategy instruction scale, 0.26 for the effective instruction behavioral scale, and 0.20 for the classroom management scale. The *ICC* estimates for the Teacher Induction study are 0.11 for the lesson content scale, 0.01 for the classroom culture scale, and 0.08 for the lesson implementation scale.

Due to this variation, a conservative mediator *ICC* value of 0.15 was assumed for the analysis, which implies an estimate of about 0.5 for ψ. This 0.5 value was also assumed for ψ^{Obs} (although ψ^{Obs} and ψ may differ in practice).

$R^2_{\bar{M}_B,T}$ *and* $R^2_{M,T}$ *values.* The $R^2_{\bar{M}_B,T}$ parameter is the population squared correlation between $\bar{M}_i$ and T_i, and is a function of the size of the treatment effect on the mediator. To obtain plausible values for this parameter, it is convenient to use the relation from (2) that $R^2_{\bar{M}_B,T} = \beta^2_{1,eff}\, p(1-p)$, where $\beta^2_{1,eff} = \beta^2_1 / \sigma^2_{\bar{M}_B}$ is the squared impact on $\bar{M}_i$ measured in *effect size* (standard deviation) units. Thus, estimates of $R^2_{\bar{M}_B,T}$ can be obtained using estimates of $\beta^2_{1,eff}$.

Two similar approaches were used for obtaining plausible values for $\beta^2_{1,eff}$. First, a "rule-of-thumb" from the IV literature is that if the $F = t^2 = \hat{\beta}^2_1 / V\hat{a}r(\hat{\beta}_1)$ statistic from (2) is 10, then T_i can be considered to be a strong instrument for $\bar{M}_i$ (see Murray 2006 and Stock et al. 2002). With 60 study schools (a typical sample size), this condition implies that $\beta_{1,eff} = 0.66$ and, thus, that $R^2_{\bar{M}_B,T} = 0.11$ (see (28) below). The second approach is to set $\beta_{1,eff}$ equal to the minimum detectable impact in effect size units (*MDE*) for the mediator. With 60 schools, this approach yields $\beta_{1,eff} = MDE = 0.51$ and $R^2_{\bar{M}_B,T} = 0.07$ (see (28) below).

Based on these analyses, an $R^2_{\bar{M}_B,T}$ value of 0.10 was used for the simulations. Importantly, this small $R^2_{\bar{M}_B,T}$ value suggests that the variance of the IV estimator will be *large*, because $R^2_{\bar{M}_B,T}$ enters the denominator of the IV variance formulas. Furthermore, this denominator term will matter unless the impact on the mediator is unrealistically large. For example, the impact on the mediator would need to be 1.4 standard deviations to yield an $R^2_{\bar{M}_B,T}$ value of 0.5, and 1.8 standard deviations to yield an $R^2_{\bar{M}_B,T}$ value of 0.8.

Finally, because $R^2_{M,T} = \psi R^2_{\bar{M}_B,T}$, an $R^2_{M,T}$ value of 0.05 was used for the simulations, which was obtained by multiplying estimates of $\psi = 0.50$ and $R^2_{\bar{M}_B,T} = 0.10$.

Empirical Results

For context, this section first presents *MDE*s for *impacts* on test score gains and a study mediator using OLS estimates of α_1 in (1) and β_1 in (2). The section then presents simulation results from the statistical power analysis for γ_1.

MDE Results

Using (1) and (2) and the methods discussed above and in Schochet (2008a), the *MDE* formulas for student test score gains and a classroom-level mediating outcome are as follows:

$$(27)\quad MDE(Test\ Score\,Gains) = 2.802\sqrt{Var(\hat{\alpha}_{1,OLS})/\sigma_y^2} = 2.802\sqrt{deff_B / ncmp(1-p)},$$

and

$$(28)\quad MDE(Mediator) = 2.802\sqrt{Var(\hat{\beta}_{1,OLS})/\sigma_M^2} = 2.802\sqrt{deff_M / \lambda ncp(1-p)},$$

where $deff_B = [1+\rho_1(m-1)+\rho_2(cm-1)]$ and $deff_M = [1+\lambda(\psi c-1)]$.

For typical RCT samples of 60 schools and 180 classrooms split evenly between the treatment and control groups and using the assumptions from above, the *MDE* on student gain scores is 0.27 (Table 5.1). With these samples, the *MDE* on a study mediator is 0.51 if $\lambda = 1$ (that is, in the absence of measurement error), 0.66 if $\lambda = 0.5$ and 0.98 if $\lambda = 0.2$ (Table 5.1). With 300 study schools, the corresponding MDEs are about half as large.

Statistical Power Results for Mediator Effects

What are likely power levels for RCT exploratory analyses that aim to estimate associations between teacher practice and student achievement measures? To help answer this question, Tables 2 to 4 present the number of schools that are required to detect targeted mediator effects with power levels (probabilities) ranging from 0.60 to 0.90. Figures are presented for mediator effects within schools, between schools, and overall. In addition, figures are presented separately for reliability values of 0.2, 0.5, and 1.0 for the mediator (as defined in equation [13a]). Table 5.2 presents figures assuming that the teacher practice mediator explains 10 percent of the variance in classroom effects, while Tables 5.3 and 5.4 assume corresponding values of 20 percent and 5 percent, respectively. Figures for the between-school mediator effects are presented for both the OLS and IV estimators.

The two main empirical findings can be summarized as follows:

Finding 1: ***For typical RCTs with about 60 total study schools, the OLS approach will yield estimates of overall and within-school mediator effects with sufficient power under two stringent conditions: (1) the reliability of the mediator must be relatively large (at least 0.50), and (2) the mediator must explain a relatively large share of the classroom-level variation in student test score gains (at least 20 percent).*** For instance, if $\lambda_{rel} = 0.5$ and the teacher practice mediator explains 20 percent of the variance in classroom effects, a statistical power level of 80 percent could be achieved with 43 schools for the overall mediator effect and 53 schools for the within-school mediator effect (middle panel of Table 5.3). Stated differently, with 43 (53) schools, the RCT would have an 80 percent probability of finding a statistically significant overall (within-school) mediator effect. In contrast, if the reliability of the mediator was instead 0.2, the numbers of required schools would be 108 and 135, respectively (bottom panel of Table 5.3). Similarly, if the mediator explains only 10 percent of the variance in classroom effects, a power level of 80 percent could only be achieved with 60 study schools if λ_{rel} was close to 1 (Table 5.2).

These two conditions are intuitive. They imply that there must be a strong association between the mediator and student gain scores (so that the mediator is capturing key dimensions of teacher practices), and that there is sufficient signal in the observed mediator (that is, high reliability) so that this strong association can be estimated precisely.

Importantly, as discussed, these conditions are stringent. The finding that the mediator must explain at least 20 percent of the variation in estimated classroom effects implies a relatively high correlation of 0.45 between the two measures. Furthermore, Raudenbush et al. (2008) demonstrate that the reliability of teacher practice measures as defined in (13a) may not be high. Thus, in practice, it is more likely that 150 to 200 schools would be required to produce precise overall and within-school mediator associations using the OLS approach (Tables 5.2 and 5.4).

Finding 2: ***For typical RCT samples, the IV approach will yield estimates with very little statistical power for detecting between-school mediator associations.*** Even in the most favorable of the considered scenarios—where $\lambda_{rel} = 1$ and the mediator explains 20 percent of the classroom-level variation in student test scores—more than 500 schools would be required under the IV approach to achieve a statistical power level of 80 percent (top panel of Table 5.3). Furthermore, more than 100 schools would be required under this best case scenario even if the impact on the mediator was 1.4 standard deviations (so that the treatment status indicator would explain about 50 percent of the variance in the mediator; not shown). Under less favorable scenarios, hundreds, or even thousands of schools would be required (Tables 5.2 to 5.4).

This low power occurs because the denominator of the asymptotic variance of the IV estimator includes the squared correlation between $\bar{M}_i$ and T_i which, as discussed, is likely to be small. Thus, although the IV approach can adjust for simultaneity and omitted variable biases that are likely to plague the OLS estimators, this approach has very little statistical power for mediator analyses.

Table 5.1: *MDE* Values for Student Gain Scores and a Teacher Practice Mediating Outcome

		***MDE* for a Teacher Practice Mediator, by Level of Reliability:**		
Number of Schools	***MDE* for Student Gain Scores**	$\lambda_{rel} = 1$	$\lambda_{rel} = 0.5$	$\lambda_{rel} = 0.2$
20	0.47	0.89	1.14	1.70
40	0.33	0.63	0.81	1.20
60	0.27	0.51	0.66	0.98
80	0.24	0.44	0.57	0.85
100	0.21	0.40	0.51	0.76
200	0.15	0.28	0.36	0.54
300	0.12	0.23	0.30	0.44

Note: See text for formulas and assumptions.

Table 5.2: Total Number of Schools Required to Detect Teacher Practice-Achievement Associations Assuming the Mediator Explains 10 Percent of the Variation in Classroom Effects, by Power Level

	Target R^2 = 0.015 for the Overall Association: γ_1 in (3)	Target R^2 = 0.005 for the Within-School Association: γ_{1W} in (4)	Target R^2 = 0.01 for the Between-School Association: γ_{1B} in (4)	
Power Level	**OLS**	**OLS**	**OLS**	**IV**
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 1$				
0.60	27	33	64	650
0.70	34	42	81	818
0.80	43	53	104	1,039
0.90	58	72	139	1,389
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 0.5$				
0.60	54	67	130	653
0.70	68	84	164	822
0.80	87	108	208	1,044
0.90	116	144	279	1,396
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 0.2$				
0.60	135	168	327	655
0.70	171	212	412	824
0.80	217	270	523	1,047
0.90	290	361	699	1,400

Note: See text for formulas and assumptions. The OLS figures were calculated using equations (18)-(20) and the IV figures were calculated using equation (26).

Table 5.3: Total Number of Schools Required to Detect Teacher Practice-Achievement Associations Assuming the Mediator Explains 20 Percent of the Variation in Classroom Effects, by Power Level

	Target R^2 = 0.03 for the Overall Association: γ_1 in (3)	Target R^2 = 0.01 for the Within-School Association: γ_{1W} in (4)	Target R^2 = 0.02 for the Between-School Association: γ_{1B} in (4)	
Power Level	**OLS**	**OLS**	**OLS**	**IV**
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 1$				
0.60	13	16	32	321
0.70	17	21	40	405
0.80	21	26	51	514
0.90	29	36	68	688
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 0.5$				
0.60	27	33	64	325
0.70	34	42	81	409
0.80	43	53	104	519
0.90	58	72	139	695
Reliability of Teacher Mediator: $\lambda_{rel} = 0.2$				
0.60	67	84	163	327
0.70	85	106	205	411
0.80	108	135	261	523
0.90	145	180	349	699

Note: See text for formulas and assumptions. The OLS figures were calculated using equations (18)-(20) and the IV figures were calculated using equation (26).

Table 5.4: Total Number of Schools Required to Detect Teacher Practice-Achievement Associations Assuming the Mediator Explains 5 Percent of the Variation in Classroom Effects, by Power Level

	Target R^2 = 0.0075 for the Overall Association: γ_1 in (3)	Target R^2 = 0.0025 for the Within-School Association: γ_{1W} in (4)	Target R^2 = 0.005 for the Between-School Association: γ_{1B} in (4)	
Power Level	**OLS**	**OLS**	**OLS**	**IV**
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 1$				
0.60	54	67	130	1,306
0.70	68	84	164	1,644
0.80	87	108	208	2,088
0.90	116	144	279	2,791
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 0.5$				
0.60	108	135	261	1,309
0.70	136	170	329	1,648
0.80	174	216	418	2,093
0.90	232	288	559	2,798
Reliability of Teacher Practice Mediator: $\lambda_{rel} = 0.2$				
0.60	272	338	655	1,311
0.70	342	425	825	1,650
0.80	434	540	1,048	2,096
0.90	581	722	1,401	2,802

Note: See text for formulas and assumptions. The OLS figures were calculated using equations (18)-(20) and the IV figures were calculated using equation (26).

Chapter 6: Summary and Conclusions

This paper has examined, both theoretically and empirically, the extent to which typical large-scale school-based RCTs in the education area will have sufficient statistical power for conducting analyses to estimate associations between teacher practice mediators and student gain scores. These exploratory analyses are of interest to quantitatively link impact estimates on teachers and students, as postulated by the study's conceptual model.

The theory in the paper developed asymptotic formulas for calculating statistical power for detecting mediator effects using two regression approaches. First, the paper considered a simple OLS (correlational) approach, which can easily accommodate multiple mediators, but which may yield biased estimates due to omitted variables, simultaneity, and measurement error. Thus, an IV approach, where treatment status is used as an instrument for the mediator, was also considered to help avoid these biases. For both approaches, the power formulas incorporate precision losses due to measurement error in the mediator.

In the empirical analysis, the theoretical formulas were used to simulate the likely statistical power of mediator analyses for the considered models. A key finding is that for typical RCTs with about 60 total study schools, OLS methods will yield precise estimates of mediator effects under two stringent conditions. First, the reliability of the observed teacher practice mediator as defined in equation (13a) must be at least 0.50. Second, the correlation between the mediator and estimated classroom effects must be at least 0.45, so that the mediator must explain a good deal of the classroom-level variation in student gain scores.

For several reasons, however, these conditions are likely to be stringent in practice. First, Raudenbush et al. (2008) demonstrate that currently available mediators from classroom observation data may have reliabilities that are lower than 0.50, due to considerable variability in rater measurements and teacher practices throughout the school day. Second, as discussed in this paper, studies of educational interventions often find weak associations between classroom practices and student outcomes, suggesting that mediator-test score correlations may be considerably lower than 0.45. Thus, it is more likely that about 150-200 schools would be required to produce precise estimates of mediator effects using the OLS approach.

The conditions under which the OLS approach will yield unbiased estimates seem unlikely to hold in practice. Thus, the IV approach may be preferable because the key condition under which it can produce unbiased estimates—the exclusion restriction that all intervention effects on student gain scores must work through the mediator—may be plausible for some interventions. However, the IV approach has very little statistical power for mediator analyses. Furthermore, there are other limitations of the IV approach, such as finding suitable instruments when multiple mediators are included in the model, the fact that only between-school mediator effects can be identified, and that full population causal effects can be estimated only under certain conditions, such as constant treatment effects.

Thus, results from this paper suggest that unless the sample contains a large number of schools (about 150-200), regression-based mediator analyses are likely to be informative only if new mediators can be developed that have higher reliabilities and stronger associations with student learning measures. Even with these improved measures, however, mediator analyses will need to rely on OLS methods—which could produce biased estimates—because sample size requirements would be prohibitively large using the IV approach.

The findings from this paper may have implications for the types of mediators that RCTs currently collect and the budget allocated to collecting these expensive data. For instance, mediators that assess the fidelity of implementation of the intervention may have descriptive importance for RCTs to help understand the impact findings. However, measures of teacher practices may be of less use if there is little chance that significant mediator-test score relationships can be detected. In these cases, the evaluation may have sufficient power for detecting impacts on the teacher practice mediators and student test scores in isolation, but would have little basis for quantitatively linking these two sets of outcomes and impacts. Thus, these classroom practice mediators may be of little help in confirming the study's conceptual model and identifying teacher practices that are most associated with student learning gains.

Appendix A: Proof of Equation (18)

The asymptotic variance of $\hat{\gamma}_{1,OLSb,ME}$ can be approximated as follows:

$$(A.1)\quad AsyVar(\hat{\gamma}_{1,OLSb,ME}) \approx \frac{\sigma^2 deff_1}{ncm\sigma^2_{M^{Obs}}(1-R^2_{M^{Obs},T})},$$

where all terms were defined in the main text. Thus, the associated non-centrality parameter is:

$$(A.2)\quad \delta_{OLSb} = \frac{(\lambda\gamma_1)^2}{AsyVar(\hat{\gamma}_{1,OLSb,ME})} \approx \frac{(\lambda\gamma_1)^2 ncm\sigma^2_{M^{Obs}}(1-R^2_{M^{Obs},T})}{\sigma^2 deff_1}$$
$$= \frac{\lambda^2\gamma_1{}^2 ncm(1/\lambda)\sigma^2_M(1-R^2_{M,T})}{\sigma^2 deff_1},$$

where the last equality holds using the definition of λ. Note that the partial squared correlation, $R^2_{y,M|T}$, can be expressed as follows:

$$(A.3)\quad R^2_{y,M|T} = \frac{\gamma_1^2\sigma^2_M(1-R^2_{M,T})}{\sigma^2_y(1-R^2_{y,T})},$$

where $R^2_{y,T}$ is the squared population correlation between y_{ijk} and T_i. Thus, solving (A.3) for $(1-R^2_{M,T})$ and inserting this expression into (A.2) yields:

$$(A.4)\quad \delta_{OLSb} = \frac{ncm(\lambda R^2_{y,M|T})\sigma^2_y(1-R^2_{y,T})}{\sigma^2 deff_1}.$$

Finally, (18) can be obtained from (A.4) using the two relations: (1) $\sigma^2 = \sigma^2_y(1-R^2_{y,M^{Obs},T})$, where $R^2_{y,M^{Obs},T}$ is the total regression R^2 value; and (2) $(1-R^2_{y,M^{Obs},T}) = (1-R^2_{y,T})(1-\lambda R^2_{y,M|T})$.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teacher and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, vol. 25, no. 1, 95-136.

Baron, R. and D. Kenny (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182.

Bloom, H., P. Zhu, and F. Unlu (2009). Finite Sample Bias from Instrumental Variables Analysis in Randomized Trials. MDRC Working Paper.

Bryk, A. and S. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Chiang (2009). Classroom and School-Level ICCs in Test Score Gains of Elementary School Students in Low Income Schools. Mathematica Policy Research Working Paper.

Cohen, J. (1977; 1988 2nd Edition). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Academic Press.

Fuller, W (1987). *Measurement Error Models.* New York: John Wiley and Sons.

Gamse, B.C., Bloom, H.S., Kemple, J.J., Jacob, R.T., (2008). Reading First Impact Study: Interim Report. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., S. Dolfin, M. Bleeker, A. Johnson, E. Isenberg, J. Lugo-Gil, M. Grider, E. Britton (2008). Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Goldhaber, D. (2002). The Mystery of Good Teaching: Surveying the Evidence on Student Achievement and Teachers' Characteristics. *Education Next*, Vol. 2, No. 1, 50-55.

Greene, W. (2000). *Econometric Analysis*. Fourth Edition. Upper Saddle River, NJ: Prentice Hall.

Hanushek, E., J. Kain, D. O'Brien, and S. Rivkin (2005). The Market for Teacher Quality. National Bureau of Economic Research Working Paper 11154.

Hedges, L. and Hedberg, E. (2007). Intraclass Correlation Values for Planning Group Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, 29, 60-87.

Holland, P.W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equation Models. In C.C. Clogg (Ed.). *Sociological Methodology* (pp. 449-493). Washington DC: American Sociological Association.

Imbens, G. and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467–476.

Jacob, B. and L. Lefgren (2005). Principals as Agents: Subjective Performance Measurement in Education. National Bureau of Economic Research Working Paper 11463.

Jackson, R., A. McCoy, C. Pistorino, A. Wilkinson, J. Burghardt, M. Clark, C. Ross, P. Schochet, and P. Swank (May 2007). National Evaluation of Early Reading First: Report to Congress. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

James-Burdumy, S. et al. (2009). Effectiveness of Selected Supplemental Reading Comprehension Interventions. Washington, DC: U.S. Institute of Education Sciences.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Kline, R (2005). *Principles and Practice of Structural Equation Modeling*. Guilford, NY: Guilford Press.

Kling, J., J. Liebman and L. Katz (2007). Experimental Analysis of Neighborhood Effects. *Econometrica* Vol. 75, No. 1, 83-119.

Kraemer, H. and S. Thiemann (1987). *How Many Subjects?* Newbury Park, CA: Sage.

MacCallum, R., M. Browne, and H. Sugawara (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. *Psychological Methods* 1(2), 130-149.

MacKinnon, D. and Dwyer, J. (1993). Estimating Mediated Effects in Prevention Studies. *Evaluation Review*, 17, 141-158.

McCaffrey, D., J. Lockwood, D. Koretz, T. Louis, and L. Hamilton (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.

Murray, M. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives*, 20(4), 111-132.

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

Perez-Johnson, I., L. Campuzano, K. Fortson, C. Gentile, S. Amin, J. Burghardt, A. Schirm (2009). Designing a Study to Validate Measures of Effective Teachers and Classrooms. Mathematica Policy Research Working Paper.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley and Sons.

Raudenbush, S., A. Martinez, H. Bloom, P. Zhu, and F. Lin (2008). The Reliability of Group-Level Measures and the Power of Group-Randomized Studies. University of Chicago Working Paper.

Raudenbush, S. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, 2, 173-185.

Rogers, W. and K. Hopkins (1988). Power Estimates in the Presence of a Covariate and Measurement Error. *Educational and Psychological Measurement*, Vol. 48, 647-656.

Rothstein, J. (2009). *Teacher Quality in Education Production: Tracking, Decay, and Student Achievement.* Princeton University Industrial Relations Section Working Paper.

Saginor, N. and P. Hyjek (2005). Observing Standards-Based Classrooms: The Vermont Classroom Observation Tool (VCOT). Montpelier, VT: Vermont Institutes.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Sobel, M. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equations Models. In S. Leinhart (Ed.), *Sociological Methodology* (pp. 290-312). San Francisco: Jossey-Bass.

Sobel, M. (2008). Identification of Causal Parameters in Randomized Studies With Mediating Variables. *Journal of Educational and Behavioral Statistics*, 33(2), 230-251.

Stock, J., J. Wright, and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics*, 20(4), 518-529.

Wooldridge, J (2002). *Econometric Analysis of Cross Section and Panel Data*. MA: MIT Press.