Designing and Analyzing Studies That Randomize Schools to Estimate Intervention Effects on Student Academic Outcomes Without Classroom-Level Information

Pei Zhu (MDRC)
Robin Jacob (University of Michigan)
Howard Bloom (MDRC)
Zeyu Xu (Urban Institute)

April 2011

Acknowledgments

The authors thank Steve Raudenbush, Andres Martinez, and Fen Lin for their contribution to the paper. The working paper was supported by William T. Grant Foundation, a staff development grant from Abt Associates Inc., and the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC, which was created through gifts from The Annie E. Casey Foundation, The Rockefeller Foundation, The Jerry Lee Foundation, The Spencer Foundation, William T. Grant Foundation, and The Grable Foundation. The authors also thank the National Center for the Analysis of Longitudinal Data in Education Research (CALDER, supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education) for providing access to some of the data used in this paper.

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

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

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

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

Abstract

This paper provides practical guidance for researchers who are designing and analyzing studies that randomize schools — which comprise three levels of clustering (students in classrooms in schools) — to measure intervention effects on student academic outcomes when information on the middle level (classrooms) is missing. This situation arises frequently in practice because many available data sets identify the schools that students attend but not the classrooms in which they are taught. Do studies conducted under these circumstances yield results that are substantially different from what they would have been if this information had been available? The paper first considers this problem in the context of planning a school-randomized study based on preexisting two-level information about how academic outcomes for students vary across schools and across students within schools (but not across classrooms in schools). The paper next considers this issue in the context of estimating intervention effects from school-randomized studies. Findings are based on empirical analyses of four multisite data sets using academic outcomes for students within classrooms within schools. The results indicate that in almost all situations one will obtain nearly identical results whether or not the classroom or middle level is omitted when designing or analyzing studies.

Table of Contents

List of Tables

Introduction

What are the implications of planning and analyzing a study that randomizes groups comprised of three levels of variation without explicitly accounting for the middle level? For example, what if one randomized schools but planned the study and analyzed the resulting data without explicitly accounting for the clustering of students within classrooms?

This problem often occurs at the planning stage of studies that randomize schools because little is known about the three-level variance structure of outcome measures for students clustered in classrooms in schools. Most of the published empirical basis for planning such studies instead comprises information for the two-level variance structure of students clustered in schools (see, for example, Bloom, Richburg-Hayes, and Black, 2007; Hedges and Hedberg, 2007). Thus, research designs based on this information do not account explicitly for the clustering of students in classrooms.

The problem also occurs at the analysis stage of studies that randomize schools because researchers often use administrative records to measure student outcomes. Because these records typically do not identify which students are in which classrooms — and adding such identifiers is difficult or costly, if not impossible to do — the resulting studies are analyzed using two-level models that do not account explicitly for the clustering of students within classrooms.

Previous researchers have considered the implications of ignoring a level of variance when analyzing data with a multilevel structure. Specifically, they have shown that if a middle level of a multilevel variance structure is ignored, part of it will shift up one level and the rest will shift down one level, thereby increasing *estimates* of the variances at these adjacent levels. In this way, the middle-level variance is to some extent accounted for implicitly (Opdenakker and Van Damme, 2000; Moerbeek, 2004; Tranmer and Steele, 2001; Van den Noortgate, Opdenakker, and Onghena, 2005).

Researchers also have tested (using simulated and actual data) the implications that such omissions can have for the interpretation of multiple regression analyses. They have demonstrated, for example, that in many situations ignoring a level of variance will result in standard errors that are misspecified and thereby produce incorrect statistical inferences. For example, omitting the classroom level in a sample that has students clustered in classrooms within schools will produce incorrect estimates of standard errors for *student-level independent variables* (Opdenakker and Van Damme, 2000; Moerbeek, 2004; Van den Noortgate, Opdenakker, and Onghena, 2005).

These studies provide a general overview of what happens to both the standard errors and point estimates of predictors included at all levels of a hierarchical model when various levels are ignored. However, as Van Landeghem, De Fraine, and Van Damme (2005) note, the findings from these studies often do not apply to situations that researchers commonly face in practice. For example, many of the results are based on the assumption that the size and internal structure of every randomized cluster is the same and that no covariates are included in the analysis. This is rarely the case in practice. Similarly, these results are usually quite general and depend on factors like the particular level of a multilevel variance structure that is ignored, the level of the predictor variable of interest, and the relative magnitudes of the variance components involved. The overarching conclusion of these papers is that omitting a level from a multilevel analysis can be problematic, but it is difficult to determine the practical implications of doing so for any given potential research application. Furthermore, these studies focus primarily on the implications of this approach for *analyzing* data when a level of variance is not explicitly acknowledged and little attention is paid to the implications of missing a level of variance for the minimum detectable effects obtained during power analyses for planning studies. Consequently, there is little practical guidance for researchers who are interested in the design and analysis of school-randomized studies when information about the classroom is not available.

This paper fills that gap by exploring the consequences of ignoring classroom-level information when *designing* or *analyzing* a school-randomized trial. It extends previous findings by investigating not only the implications of not acknowledging the middle level for analyzing data, but also by investigating its implications for planning studies with a three-level data structure using only the top and bottom levels of information. The paper also provides concrete guidance to education researchers who are designing and analyzing data from school-level random assignment studies in which the cluster size and structure varies and covariates are used for analysis. Finally, the paper extends the findings to cases in which the sample used to plan an impact study has a different cluster structure (that is, a different number of students per classroom and classrooms per school) than the structure of the sample for the impact study itself. These extensions are based on empirical analyses of four multisite data sets that use academic outcomes for students within classrooms within schools.

The resulting findings indicate that no substantial problem is likely to arise from using two-level models (for students within schools) to design or analyze studies that randomize schools. This conclusion holds for both elementary school data (where the middle-level variance component tends to be small) and secondary school data (where the middle-level variance component tends to be large), for data sets with varying numbers of students per classroom and classrooms per school, in situations where covariates are

included at either the student or school level, and in situations where the cluster structure of the study being planned differs substantially from the one used for planning purposes.

The rest of the paper is structured as follows. It begins by presenting a theoretical framework for comparing three- and two- level models of a three-level situation. The paper then presents estimates of three-level and two-level variance components and examines how an ignored classroom-level variance component is shifted up to the school level and down to the student level. The authors compare the shifting in their data with what is predicted theoretically and find the actual and predicted shifts to be consistent with each other. These findings are then used to consider the implications of a two-level analysis of minimum detectable effect sizes (MDES) for a study that randomizes schools. These implications are explored for models that do and do not use covariates to estimate MDES. The paper also explores the implications of planning a study that has a different underlying data structure than the one used for planning purposes. The paper next explores the implications of ignoring the middle level when analyzing data with a three-level structure using a two-level model, and it ends by offering some conclusions and recommendations.

Theoretical Framework

Consider the following two alternative research designs for estimating the impacts of an educational intervention on student outcomes from a study that randomizes schools in a large urban district. Both designs will estimate impacts by the observed differences in mean student outcomes for the randomized treatment group and control group, and the true variance structure for the study's sample will comprise three levels: students, classrooms, and schools.

Design A uses a statistical model that specifies all three levels of the true variance structure. The school-level variance equals τ_A^2, which is the variance of mean outcomes across schools within the district. The classroom-level variance equals γ_A^2 , which is the variance of classroom means within schools. The student-level variance equals σ_A^2, which is the variance of student scores within classrooms. The total student variance equals the sum of these three variance components ($\tau_A^2 + \gamma_A^2 + \sigma_A^2$).

Design B uses a two-level statistical model that specifies two variance components, one for mean values of the outcome measure across schools, τ_B^2, another for individual student outcomes within schools, σ_B^2. These two variances sum to the total student variance, which is the same as that for the three-level model but is decomposed differently. Because the clustering of students within classrooms is

ignored, student outcomes are assumed to vary independently of each other within schools, which is an oversimplification.

The following expressions can be used to compute a minimum detectable effect size for student outcomes given designs A and B, without covariates or blocking. Note that throughout this paper, minimum detectable effect sizes are defined for a two-tail hypothesis test at the 0.05 level of statistical significance with 80 percent statistical power (for a discussion of how this is done, see Bloom, 2005).

Design A

$$MDES_A = \frac{M_{J-2}}{\sqrt{P(1-P)}}\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}} * \frac{1}{\sqrt{\tau_A^2 + \gamma_A^2 + \sigma_A^2}} \quad (1)$$

where

$MDES_A$ = the minimum detectable effect size for design A;

M_{J-2} = a multiplier for J-2 degrees of freedom that equals approximately 2.8 for studies that randomize 20 or more schools;

P = the proportion of schools randomized to treatment;

J = the total number of schools randomized to treatment or control status;

K = the harmonic mean number of classrooms per school;

N_A= the harmonic mean number of students per classroom.

Design B

$$MDES_B = \frac{M_{J-2}}{\sqrt{P(1-P)}}\sqrt{\frac{\tau_B^2}{J} + \frac{\sigma_B^2}{JN_B}} * \frac{1}{\sqrt{\tau_B^2 + \sigma_B^2}} \quad (2)$$

Where, in addition:

N_B = the harmonic mean number of students per school.

These two expressions are the same with respect to the multiplier (M_{J-2}), which converts standard errors of estimates to minimum detectable effects (see Bloom, 2005, for a discussion). The two expressions are also the same with respect to the proportional allocation of randomized groups to treatment

status (P) and control status (1-P). However, they differ with respect to the square root of the sum of variance contributions from the different levels of each statistical model.

The central question to address when comparing these two expressions is: How do their *estimated* values compare when the total student variance is decomposed into all three components (for schools, classrooms, and students — as in Equation 1 — to when the total student variance is decomposed only into components for schools and for students within schools (as in Equation 2)?

To understand this question, first recall that both models start with the same total variance in the outcome measure across all students from all classrooms in all schools. Hence, the sum of the three variances under model A equals the sum of the two variances under model B or:

$$\tau_A^2 + \gamma_A^2 + \sigma_A^2 = \tau_B^2 + \sigma_B^2 \tag{3}$$

Variance estimates for model B must thus shift the true middle-level variance to the bottom level, the top level, or both levels.

Moerbeek (2004, Equation 14) derives the following expressions that represent this shifting when there is a constant number of classrooms per school (K) and students per classroom (N_A).

$$E(\hat{\tau}_B^2) = \tau_A^2 + \left(\frac{N_A - 1}{N_A K - 1}\right)\gamma_A^2 \tag{4}$$

$$E(\hat{\sigma}_B^2) = \sigma_A^2 + \left(\frac{N_A (K - 1)}{N_A K - 1}\right)\gamma_A^2 \tag{5}$$

Where

$E(\hat{\tau}_B^2) =$ the expected value of $\hat{\tau}_B^2$, and

$E(\hat{\sigma}_B^2) =$ the expected value of $\hat{\sigma}_B^2$.

Equations 4 and 5 indicate that a predictable portion of the true classroom-level variance is shifted to the estimated school-level variance, and the remainder is shifted to the estimated student-level variance. The sum of these two increments equals the total classroom variance.

Intuitively, it is easy to see how part of the true classroom-level variance shifts down to the estimated student-level variance. This occurs because part of the observed variance in outcomes across students within schools reflects classroom differences. Thus, when the variation across students within schools is measured and when cross-classroom differences are ignored, a part of these differences is included in the measure of student-level variance within schools, $\hat{\sigma}^2_B$. Consequently, the estimated student-level variance within schools for the two-level model $\hat{\sigma}^2_B$ exceeds that for the estimated student-level variance within classroom in the corresponding three-level model, $\hat{\sigma}^2_A$.

It is less readily apparent how the two-level estimation model B attributes some of the cross-classroom variance to the estimated variance across schools. This occurs because model B assumes that outcomes vary independently across students within schools, when in fact they are clustered by classroom. By ignoring the clustering of students within classrooms, the two-level model B *understates* the contribution of student-level variation to the total observed variance of school sample *means*. Equation 4 indicates that more of the classroom-level variance is shifted to the estimated school-level variance as students per school (N_AK) are clustered into fewer classrooms (K). This shift reflects how the clustering of students within classrooms inflates the true variability of within-school outcomes. Ignoring this clustering thus causes larger understatement of the within-school variability of outcomes when there are fewer classroom clusters, which, in turn, causes one to overstate the between-school variance accordingly. In other words, when decomposing the total observed variance in school sample means into variation due to true variation across schools and variation due to estimation error produced by within-school student variation, the two-level model *overestimates* the true school-level variance. Consequently, the estimated school-level variance for the two-level model exceeds that for the three-level model.

Because the classroom variance that is ignored by a two-level model is *reflected* in estimates of school and student variances, the classroom variance *is not missing* from a two-level analysis. Indeed, as has been shown by others (Moerbeek, 2004; Van den Noortgate, Opdenakker, and Onghena, 2005) theoretically, using a two-level model to estimate the cross-level variance components to be used in the calculation of the minimum detectable effect for a group-randomized research design will produce the same results as those produced by a three-level model. As noted, however, these theoretical conclusions assume that every school has the same number of classrooms per school and students per classroom, that data used for planning a study reflect the number of classrooms per school and students per classroom that will be included in the actual study sample, and that no covariates will be used for the study's

analysis. To extend these theoretical findings to situations that occur in practice, the remainder of this paper explores empirically what happens when the middle level of a three-level model is excluded from analyses, using three-level student outcome data from four major sources.

The Data

Data from four different sources are used for the following analysis. They are the School Breakfast Pilot Project (Abt Associates Inc. and Promar International, 2005), the federal Reading First Impact Study (Gamse, Bloom, Kemple, and Jacob, 2008) and statewide administrative-records data on standardized test scores for individual students in multiple subjects from Florida and from North Carolina. Tables 1 and 2 describe the size, structure, and variability of the analysis samples for each data source. As can be seen, these analysis samples provide an unusually large, diverse, and comprehensive empirical basis of analysis.

Table 1 reports the numbers of districts represented by these data plus the harmonic mean numbers of schools per district, classrooms per school, and students per classroom in the analysis sample. Of particular importance is the fact that the internal cluster structure of schools in the sample (that is, their number of classrooms and students per classroom) varies widely *across* the four data sources. Because, as demonstrated by Equations 4 and 5, this internal cluster determines how the classroom-level variance is shifted upward (to the school level) and downward (to the student level) when the middle level is ignored, it is important to represent a wide range of cluster structures in the analysis.

Table 2 describes the variability *within* the sample from each data source in its number of schools per district, number of classrooms per school, and number of students per classroom. This variability is measured by the standard deviation of each parameter. Of particular importance is the substantial variability that exists in the internal cluster structure of schools (their numbers of classrooms and students per classroom). This variability is what enables this paper to extend past theoretical work in ways that provide practical guidance for designing and analyzing educational evaluations (recall that existing theoretical findings *assume no such variability*).

The School Breakfast Pilot Project (SBPP) was a three-year demonstration (2000-03) that used a matched-pair random-assignment design to randomly assign schools within six districts to a treatment condition in which schools implemented a universal free school-breakfast program or to a control

Table 1. Data Structure for Each Study and Outcome

Outcome	Number of Districts	Harmonic Mean Number of Schools per District	Harmonic Mean Number of Classes per School (K)	Harmonic Mean Number of Students per Class (N)
SBPP:				
Stanford 9 Total Math scaled score	4	7.94	2.06	3.72
Stanford 9 Total Reading scaled score	4	5.86	2.05	3.74
RFIS:				
SAT 10 reading comprehension test grade 1	16	9.49	3.17	9.21
SAT 10 reading comprehension test grade 2	16	9.58	3.11	9.16
SAT 10 reading comprehension test grade 3	15	9.79	3.01	9.69
FL Elementary School Data:				
FCAT Math test for grade 5	43	6.34	4.13	17.30
FCAT Reading test for grade 5	43	6.34	4.13	17.27
NC Elementary School Data:				
Math test for grade 5	86	5.26	2.94	16.33
Reading test for grade 5	86	5.26	2.94	16.30
NC Secondary School Data:				
High School Algebra 2	41	2.99	3.17	19.75
High School Biology	48	2.91	3.00	15.90
High School Chemistry	29	2.63	2.82	18.38
High School Geometry	44	2.96	2.97	19.40
High School Physics	6	2.86	2.64	15.39

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

Table 2. Variation in Data Structure for Each Study and Outcome

Outcome	Standard Deviation of Number of Schools per District	Standard Deviation of Number of Classes per School	Standard Deviation of Number of Students per Class
SBPP:			
Stanford 9 Total Math scaled score	18.97	0.28	1.27
Stanford 9 Total Reading scaled score	18.95	0.26	1.25
RFIS:			
SAT 10 reading comprehension test grade 1	7.01	1.44	4.75
SAT 10 reading comprehension test grade 2	6.57	1.56	4.77
SAT 10 reading comprehension test grade 3	6.85	1.45	4.56
FL Elementary School Data:			
FCAT Math test for grade 5	43.44	1.91	4.75
FCAT Reading test for grade 5	43.44	1.91	4.76
NC Elementary School Data:			
Math test for grade 5	11.77	1.19	3.86
Reading test for grade 5	11.77	1.19	3.86
NC Secondary School Data:			
High School Algebra 2	3.33	2.32	4.95
High School Biology	2.96	2.04	5.58
High School Chemistry	2.55	2.50	4.95
High School Geometry	5.24	2.11	5.28
High School Physics	3.25	1.00	4.84

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

condition in which schools continued to operate their regular subsidized breakfast programs for eligible students from low-income families. The goal of the project was to measure the added value of universal free school breakfasts. The two outcome measures used from the SBPP for the analysis presented in this paper are the Stanford 9 Total Math Scale Score (math achievement test scores in scaled-score points) and the Stanford 9 Total Reading Scale Score (reading achievement test scores in scaled-score points).

The present SBPP analysis sample contains 1,151 third-graders from 233 classrooms in 111 schools from six districts. On average, there are approximately 3.7 students per classroom and two classrooms per school. The number of schools per district varies from around six to eight depending on the outcome measure (see Table 1). The cluster structure of the SBPP sample is relatively constant *by design* because the original study sampled a fixed number of classrooms per school and students per classrooms. Hence, this sample has the smallest standard deviations for these parameters (see Table 2).

The Reading First Impact Study was a three-year (2004-07), congressionally mandated evaluation of the federal government's Reading First initiative to help all children read at or above grade level by the end of third grade (Gamse, Bloom, Kemple, and Jacob, 2008). The study used a regression discontinuity design that capitalized on the systematic process used by some districts to allocate their Reading First funds to schools. The study was designed to measure the effects of the program on teacher practices and student achievement. Seventeen districts plus one state program were chosen for the study, and its original sample included 248 schools. Data for the present analysis are limited to 15 sites (14 districts plus one state) and 225 schools for which it was possible to estimate student, classroom, and school-variance components. Reading First outcome measures used for the present analysis are SAT 10 reading scaled scores for all first, second, and third-graders in the study's schools during the spring of 2005.

Even though the RFIS was a regression discontinuity analysis, it was possible to use its data to explore the implications of these data for a research design that would have randomized the schools. This was accomplished by ignoring the rating variable used to allocate Reading First funds (which was the basis for the study's regression discontinuity analysis) and estimating the natural variation in academic outcomes that exists across schools, classrooms within schools, and students within classrooms.

The RFIS sample for the analysis presented in this paper includes approximately 10 schools per district, three classrooms per school, and nine students per classroom. Unlike the SBPP, the RFIS was not designed to have a constant cluster size and structure. Instead, all first- through third-grade students in

regular education classrooms in the study's schools were included in its original sample. Hence, there is more variability across RFIS schools in the number of students per classroom and classrooms per school than is the case for SBPP schools.

Statewide data on test scores for individual students from Florida were obtained from the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW). The FL-EDW is a longitudinal data system that includes records on all students, teachers, and schools in the state. Each year Florida students in the third through eleventh grades take the Florida Comprehensive Assessment (FCAT-SSS) in reading and math. The analysis presented in this paper uses data on these test scores for grade five (representing elementary school) in math and in reading for school year 2005-06. All scores are normalized by subject. Samples are limited to students with valid test scores in both the current year and the previous year. The analytic samples are further restricted to self-contained classrooms only. On average, this elementary school sample comprises approximately 17 students per class, four classrooms per school, and six schools per district from a total of 43 districts.

Statewide data on test scores for individual students from North Carolina were obtained from the North Carolina Education Research Data Center (NCERDC) for end-of-course assessments in reading and mathematics given to students in grades three through eight in school year 2005-06. The present analysis uses fifth-grade scores to represent scores for elementary schools. Similar to what has been done with the Florida data, the analysis keeps students with valid test scores in both the current and the previous years and it keeps self-contained classrooms only. On average, the elementary school sample has about 16 students per classroom, three classrooms per school, and five schools per district. A total of 86 districts are included in the elementary school sample.

Scores are also available for North Carolina secondary school students' end-of-course assessments in algebra II, biology, chemistry, and geometry courses in school year 2005-06. This paper uses these scores to represent scores for secondary school. These end-of-course tests allow straightforward assignment of students to classrooms. The disadvantage of having end-of-course tests, on the other hand, is that students take these tests only once and therefore no repeated measures of student performance in a particular subject are available. Thus, in order to control for pretest scores in the models that will be presented, students' test scores on algebra I are used to approximate their starting levels. On average, the secondary school sample has approximately 15 to 20 students per classroom, three classrooms per school, and three schools per district. These data represent between six and 48 districts depending on the test subject.

Because the Florida and North Carolina data are for entire states, they reflect substantial variation in the number of students per classroom and classrooms per school. Hence, as can be seen from the standard deviations of the number of classrooms per school and the number of students per classroom reported in Table 2, the data exemplify schools with varying internal cluster structures. In order to investigate the impact of ignoring the middle level (the classroom level) in the context of estimating intervention effects, half of the schools in Florida and North Carolina were randomly assigned to the "treatment" group and the other half to the "control" group, such that the true "intervention effects" should be zero.

Estimating Variance Components from Each Data Source

This section reports estimated variance components from each of the preceding data sources. Three-level variance components for design A were estimated using a three-level hierarchical linear model (student-class-school); two-level variance components for design B were estimated using a two-level hierarchical linear model (student-school). To reflect the typical range of common practices in educational evaluation research, for each design, these variance components were estimated separately for models without covariates and for models with school-level or student-level baseline test scores as a covariate. Models for the SBPP and RFIS samples include a zero/one indicator variable to distinguish between treatment schools and control schools. This was not necessary for the Florida and North Carolina samples because they do not comprise a specific set of treatment and control schools. To ensure that all analyses are based solely on variation within school districts, zero/one indicator variables for each district are included in the model. This is equivalent to centering all variables on the mean values for their blocks (see Wooldridge, 2002).

Estimated Variance Components

Table 3 presents estimated variance components for all the outcomes in the data sets used in this paper. The first three columns of Table 3 report estimated variance components for the three levels (school-class-student) of model A, and the last two columns report estimated variance components for the two levels (school-student) of model B. Each estimated variance component is standardized and reported as a proportion of the total student-level variance for the sample that it represents. Values for the three variance components in model A sum to one, and values for the two variance components in model B sum to one. Hence, the standardized variance components for schools and classrooms in these models

Table 3. Three-Level vs. Two-Level Model Comparisons: Unconditional Intraclass Correlations (ICC) at Various Levels

	Unconditional ICC				
	3-Level Model			2-Level Model	
Outcome	School-Level	Class-Level	Student-Level	School-Level	Student-Level
SBPP:					
Stanford 9 Total Math scaled score	0.085	0.029	0.886	0.097	0.903
Stanford 9 Total Reading scaled score	0.064	0.070	0.865	0.094	0.906
RFIS:					
SAT 10 reading comprehension test grade 1	0.073	0.063	0.863	0.095	0.905
SAT 10 reading comprehension test grade 2	0.043	0.066	0.892	0.060	0.940
SAT 10 reading comprehension test grade 3	0.039	0.073	0.888	0.061	0.939
FL Elementary School Data:					
FCAT Math test for grade 5	0.100	0.140	0.760	0.132	0.868
FCAT Reading test for grade 5	0.082	0.123	0.795	0.109	0.891
NC Elementary School Data:					
Math test for grade 5	0.088	0.094	0.818	0.118	0.882
Reading test for grade 5	0.071	0.063	0.866	0.090	0.910
NC Secondary School Data:					
High School Algebra 2	0.124	0.376	0.500	0.222	0.778
High School Biology	0.077	0.293	0.630	0.159	0.841
High School Chemistry	0.072	0.295	0.632	0.154	0.846
High School Geometry	0.158	0.356	0.487	0.276	0.724
High School Physics	0.165	0.342	0.493	0.259	0.741

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated values for the intraclass correlations were obtained from a three-level model and a two-level model of the outcome measure without covariates. All analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

represent intraclass correlations (that is, the proportion of total student variation that is at the school and at the classroom level, respectively). In addition, what is not shown in the table but was documented empirically is that, in all cases, the sum of the estimated nonstandardized three-level variance components equals the sum of the estimated nonstandardized two-level variance components (as noted by Equation 3).

For SBPP Stanford 9 Math scores (represented by the first row in Table 3), the standardized variance for schools in the three-level analysis equals .085. This means that 0.085 (or 8.5 percent) of the total variation across students in the analysis sample (within district blocks) is estimated to reflect differences in mean outcomes across schools. In other words, the school-level intraclass correlation equals .085. The standardized variance for classrooms in the three-level analysis equals 0.029. This means that 0.029 (or 2.9 percent) of the total variation across students in the analysis sample (within district blocks) is estimated to reflect differences in mean outcomes across classrooms within schools. In other words, the classroom-level intraclass correlation equals 0.029. The remaining proportion of total student variation (0.886) is due to differences in outcomes for students within classrooms. If instead of using a three-level model, variance components for the same data are estimated ignoring the classroom level, the estimated school-level intraclass correlation is 0.097 and that for students within schools is 0.903 (see columns 4 and 5 in Table 3).

The important point to note about these findings is that the classroom-level variance in the three-level model is shifted both to the school-level variance and to the student-level variance in the two-level model. Specifically, the estimated school-level variance for the two-level model (0.097) is larger than that for the three-level model (0.085), and the estimated variance for students within schools in the two-level model (0.903) is larger than that for students within classrooms in the three-level model (0.886). These differences are quite small, however, because the estimated classroom-level variance (0.029) is only a small proportion of total student variation. These differences, and the degree of "level shifting" they represent, are more pronounced for other samples in the table that have a greater proportion of their variation at the classroom level.

Note that the estimated classroom variance for elementary schools is consistently a much smaller proportion of total student variation than it is for secondary schools. For elementary schools this proportion is always below 0.140 — and in most cases is well below this value. In contrast, for secondary

schools the proportion ranges from 0.293 to 0.376.[1] This striking difference probably reflects more extensive student tracking in secondary schools than in elementary schools.

Comparing Predicted Versus Actual Shifting of the Middle-Level Variance Component

The empirical findings presented in Table 3 are consistent *in direction* with Equations 4 and 5, which predict the upward and downward shifting of an ignored classroom variance component. However, as already noted, Equations 4 and 5 assume a constant number of classrooms per school and students per classroom, whereas the samples used to estimate variance components for the analysis presented in this paper (and for almost all others in education research) comprise schools that vary in these regards. Table 4 thus assesses the extent to which this variation in the internal structure of schools (clusters) causes the actual shifting in the classroom-level variance to differ from the amount of shifting predicted by Equations 4 and 5.

The first two columns in the table report the actual percentage of the classroom-level variance that is shifted to the school level and student level respectively, and the last two columns present the corresponding percentages that are predicted by Equations 4 and 5 based on the harmonic mean values for the number of classrooms per school and students per classroom in the analysis sample for each data source. Even though there was variability in the underlying cluster structure of the various data sets that were explored by this analysis, the distribution of the classroom-level variance to the school and student levels is fairly consistent with what the formula predicts.

It is also worth noting that even in situations where the *percentage* of variation shifted to each level differs from the theoretical prediction, the difference between the predicted and *actual amount* of variance that is shifted to each level may still be small if the middle-level variance component was small to begin with, as is the case with most of the elementary school data used in this analysis.

[1]This pattern is also observed in the tenth-grade reading and math FCAT score for Florida secondary schools. The class-level ICC for the FCAT math test score for grade 10 is 0.486 and for the FCAT reading test score for grade 10 the class-level ICC is 0.348. Note, however, that at the secondary school level in Florida, math and English language/arts courses are much more diversified (with over 50 math-related courses and over 80 English language-related courses to choose from). Most students also take more than one such course in a year. In order to select a classroom for each student that best corresponds to the end-of-grade math test or the end-of-grade reading test, the courses taken by a student were ranked by how frequently those courses were taken by tenth-grade students, and the student's classroom was defined by the most frequently taken course. Because this classroom-assignment approach is rather arbitrary, these results were not included in the main discussion of the paper but rather were used as references.

Table 4. Three-Level vs. Two-Level Model Comparisons: Percentage of Variance Shifted to School and Student Level When Class Level Is Not Explicitly Accounted For

	Actual		Predicted by Forumla	
Outcome	% Class-Level Variance Shifted to School-Level	% Class-Level Variance Shifted to Student-Level	% Class-Level Variance Shifted to School-Level	% Class-Level Variance Shifted to Student-Level
SBPP:				
Stanford 9 Total Math scaled score	42.90	57.10	40.88	59.12
Stanford 9 Total Reading scaled score	42.86	57.14	41.10	58.90
RFIS:				
SAT 10 reading comprehension test grade 1	33.94	66.06	29.13	70.87
SAT 10 reading comprehension test grade 2	26.37	73.63	29.72	70.28
SAT 10 reading comprehension test grade 3	29.89	70.11	30.83	69.17
FL Elementary School Data:				
FCAT Math test for grade 5	22.93	77.07	23.13	76.87
FCAT Reading test for grade 5	22.12	77.88	23.13	76.87
NC Elementary School Data:				
Math test for grade 5	31.79	68.21	32.64	67.36
Reading test for grade 5	30.87	69.13	32.64	67.36
NC Secondary School Data:				
High School Algebra 2	26.06	73.94	30.44	69.56
High School Biology	28.13	71.87	31.90	68.10
High School Chemistry	27.49	72.51	34.13	65.87
High School Geometry	33.19	66.81	32.47	67.53
High School Physics	27.47	72.53	36.36	63.64

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

Planning a Study: Estimated Minimum Detectable Effect Sizes

Given the estimated variance components based on the three-level and two-level models, the next step in this analysis was to explore how not explicitly acknowledging the middle level affects the actual estimates of minimum detectable effect size for each of the outcomes in the data — this indicates the predicted level of precision one could expect to obtain for a study with a given sample size. The analysis began by estimating the precision of three-level and two-level analyses for a planned study with a cluster structure that is identical to the one from which the multilevel variances were estimated. For example, for the SBPP it is assumed that the study being planned would include approximately two classrooms per school and approximately four students per classroom (see Table 1). These findings do not necessarily extrapolate to the typical situation in practice, where multilevel variances are computed from data for an existing study and then used to design a future study with a different sample size and structure. This situation will be explored later.

The findings from this analysis are presented in Table 5. The analysis uses the standardized variance estimates from Table 3 plus the harmonic mean number of students per classroom and classrooms per school (as shown in Table 1) for each outcome measure to compute the minimum detectable effect size for that measure given its original sample structure (within a school). The total number of schools was assumed be 60, and there were assumed to be 30 treatment schools and 30 control schools for all outcomes. Equation 1 was used to compute minimum detectable effect sizes for three-level analyses and Equation 2 was used for two-level analyses.

The first set of columns in the table shows findings from models that do not include any covariates (other than treatment indicators and district indicator variables where applicable). The first column presents the minimum detectable effect size for the three-level model for each measure, and the second column shows the difference between the minimum detectable effect size for the three-level model and the corresponding two-level model (the three-level estimate minus the two-level estimate). Consider yet again the findings for the SBPP Stanford 9 math score. Assuming 60 schools with 2.06 classrooms per school and 3.72 students per class (from row 1, Table 1) plus the three-level standardized unconditional variance estimates of 0.085, 0.029, and 0.886 for schools, classrooms, and students, respectively (from row 1, Table 3), an unconditional minimum detectable effect size of 0.341 was computed using Equation 1. Similarly, assuming 60 schools and an average of 7.67 students per school (2.06 classroom x 3.72 students) and the two-level standardized unconditional variance estimates of 0.097

Table 5. Difference Between Three-Level vs. Two-Level Model Comparisons: Minimum Detectable Effect Sizes (MDES), Orginal Sample Structure

	Minimum Detectable Effect Size (# of schools = 60, T/C = 1:1)					
	No Covariates		School-Level Pretest		Student-Level Pretest	
Outcome	3-Level Model	Difference (3 lvl-2 lvl)	3-Level Model	Difference (3 lvl-2 lvl)	3-Level Model	Difference (3 lvl-2 lvl)
SBPP:						
Stanford 9 Total Math scaled score	0.341	0.000	0.256	0.003	0.316	-0.001
Stanford 9 Total Reading scaled score	0.339	-0.001	0.254	0.005	0.333	0.000
RFIS:						
SAT 10 reading comprehension test grade 1	0.258	-0.003	0.197	-0.003	0.216	-0.005
SAT 10 reading comprehension test grade 2	0.227	0.003	0.164	0.015	0.186	0.001
SAT 10 reading comprehension test grade 3	0.226	0.001	0.171	0.005	0.192	0.000
FL Elementary School Data:						
FCAT Math test for grade 5	0.280	0.000	0.183	0.005	0.155	0.000
FCAT Reading test for grade 5	0.258	0.001	0.166	0.007	0.136	0.000
NC Elementary School Data:						
Math test for grade 5	0.272	0.001	0.214	0.003	0.191	-0.001
Reading test for grade 5	0.245	0.001	0.190	0.003	0.158	0.000
NC Secondary School Data:						
High School Algebra 2	0.368	0.012	0.319	0.022	0.247	0.013
High School Biology	0.319	0.009	0.271	0.009	0.231	-0.001
High School Chemistry	0.320	0.017	0.287	0.028	0.231	0.000
High School Geometry	0.393	-0.002	0.281	0.032	0.237	0.009
High School Physics	0.408	0.020	0.344	0.024	0.273	0.026

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated values for the intraclass correlations were obtained from a three-level model and a two-level model of the outcome measure without covariates. A school-level pretest and a student-level pretest measure were used in the model to obtain the R-squared values used in the MDES calculation for models with covariates. In addition, all analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

and 0.903 (from row 1, Table 3), an unconditional minimum detectable effect size of 0.342 was computed using Equation 2. The difference between these two minimum detectable effect sizes (0.000) is shown in the second column of Table 5.

Results show that for the elementary school data, where the classroom-level variance is relatively small, the predicted level of precision is essentially the same, whether the study was planned using a two-level analysis or a three-level analysis. For the elementary school data, estimates of minimum detectable effect sizes from the two- and three-level models differ by less than .005 in all cases. While the differences in estimates of minimum detectable effects sizes are slightly larger among the secondary school data, where the classroom-level variances components were substantially larger (ranging from 0.293 to 0.376), from a substantive perspective they remain quite small in absolute terms. So in the data sets explored in this analysis, if one did not explicitly acknowledge the middle level of clustering in designing a study with a data structure that was identical to the one used for planning purposes, one would, at worst, overstate the minimum detectable effect size by 0.021 for the North Carolina physics exam, which is about a 5 percent difference in precision. From a substantive perspective this is a small difference.

Including Covariates

The findings shown in Table 5 also move comparisons of two- and three-level analyses one step further by taking the inclusion of covariates into account. In practice, baseline characteristics such as students' prior test scores and demographics are often used as covariates to improve the precision of impact estimates; yet theoretical explorations of the implications of not explicitly acknowledging the middle level assume that no covariates are included. Therefore, to see how the inclusion of covariates would influence the results shown in the first column of Table 5, the researchers conducted a corresponding set of analyses in which either a school-level pretest variable (second set of columns) or a student-level pretest variable (third set of columns) was included.

To the extent that covariates predict the variation in outcomes across individuals, classrooms, or schools, they reduce the "unexplained" variance at each of these levels. This, in turn, reduces the standard error of the impact estimate. Therefore, with covariates, the formula for computing the minimum detectable effect size for a three-level model (Model A) becomes:

$$MDES_A = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau_A^{\,2}(1-R^2_{sc})}{J} + \frac{\gamma_A^{\,2}(1-R^2_{cl})}{J*K} + \frac{\sigma_A^{\,2}(1-R^2_{st})}{J*K*N_A}} * \frac{1}{\sqrt{\tau_A^{\,2} + \gamma_A^{\,2} + \sigma_A^{\,2}}} \quad (6)$$

And for a two-level model (Model B) the formula becomes:

$$MDES_B = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau_B{}^2(1-R^2_{sc})}{J} + \frac{\sigma_B{}^2(1-R^2_{st})}{J*K*N_B}} * \frac{1}{\sqrt{\tau_B{}^2 + \sigma_B{}^2}} \quad (7)$$

Where R^2_{sc} = the explanatory power of covariates for outcome differences between schools;

R^2_{cl} = the explanatory power of covariates for outcome differences between classrooms within schools;

R^2_{st} = the explanatory power of covariates for outcome differences across students within classrooms; and

C = the number of school-level covariates in the model.

All other parameters are defined as in Equation 1 and 2.

The R-squared values are calculated as the proportion of each unconditional variance that is explained by the covariates, that is, for level L, where L = school, classroom, or student,

$$R^2_L = \frac{\sigma^2_{U,L} - \sigma^2_{C,L}}{\sigma^2_{U,L}} \quad (8)$$

Where $\sigma^2_{U,L}$ is the unconditional variance at level L when no covariates are included in the model, $\sigma^2_{C,L}$ is the conditional variance at level L when covariates are added.

Based on these estimated R-squared values (presented in Appendix Table A-1) and the original unconditional variances presented in Table 3, it is possible to use Equations 6 and 7 to estimate the minimum detectable effect size for the original sample given available covariates. To do so for the school-level pretest, the researchers included the R-squared values obtained after including a school-level pretest in Design A and substituted them in Equation 6 above. For the student-level pretest, the researchers included the R-squared values obtained for school, classroom, and student levels after including a student-level pretest in Design A. In all cases, the unconditional variances and total number of students, classrooms, and schools remained the same as in previous models.

The findings from these analyses are presented in the second and third sets of columns in Table 5. The first point to notice about these results is that including a pretest as a covariate at either at the school or student level causes an overall reduction in the minimum detectable effect size (a finding that is consistent with prior research). Take again the SBPP math score. Without covariates, the MDES for the three-level analysis is 0.341. With a school-level pretest variable, the MDES from the three-level analysis

is reduced to 0.256 and with a student-level pretest, the MDES from the three-level analysis is reduced to 0.316. A similar reduction is seen in the two-level models.

The second point to notice about the results presented in the second and third set of columns in Table 5 is that including a school-level covariate in the models used to estimate the MDES tends to exacerbate the difference between the predicted MDES obtained from a three-level analysis and the comparable two-level analysis relative to models that included no covariates. In all instances, the difference between the three-level and two-level estimate of MDES is larger than in the unconditional model. Furthermore, in all cases the minimum detectable effect size would be underestimated if a two-level model were used. This is because including the pretest at the school level reduces the variance at the school level, thereby increasing the relative amount of variance that is accounted for at the classroom level. In other words, there is relatively more classroom-level variance that could potentially be shifted to the school level when the middle level is not acknowledged. However, the differences between the two- and three-level analyses remain quite small, especially for the elementary school data. The largest difference is 0.032, for the North Carolina high school geometry test. Thus, although the inclusion of a school-level pretest makes the difference between the two- and three-level analyses larger, in no case does one observe a distortion that is substantively important.

On the other hand, as can be seen in the third set of columns, the inclusion of a student-level pretest variable reduces the difference between the estimated MDES obtained from the two- and three-level analyses. In all instances, the differences between the predicted MDES from the three- and two-level analyses are smaller when the student-level pretest variable is included than is the case for the unconditional analyses. Additionally, including a student-level pretest seems to eliminate some of the largest differences that were observed in the unconditional analyses. With the inclusion of the student-level pretest variable, for example, the difference between the two- and three-level analyses for the North Carolina high school chemistry test is reduced from 0.017 to 0.000. The largest difference between the predicted MDES from the three- and two-level analyses is 0.026 (North Carolina physics) when a student-level pretest is included. Note that algebra I scores were used as pretest measures for North Carolina secondary school subjects. This large difference may reflect the fact that algebra I is not a very good proxy for students' previous knowledge of physics. It is not hard to see why including the student-level pretest helps to reduce the problem. As shown earlier, the problem in the secondary school data is being driven by the large classroom-level variance component. If a student-level pretest is included, this classroom-level variance component is reduced substantially because much of it is accounted for by the student-level covariate. On the other hand, including a school-level pretest tends to exacerbate the

problem because the school-level pretest only reduces variance at the school level, making the relative size of classroom-level variance even bigger.

In summary, these findings illustrate that minimum detectable effect sizes computed from a two-level analysis, even when school-level or student-level covariates are included, are quite similar to those computed from a three-level analysis with the same data and covariates.

Varying the Sample Structure

The findings presented in Table 5 do not necessarily extrapolate to the typical situation in practice where multilevel variances are computed from data for an existing study and are then used to design a future study with a different sample structure. One way to emulate this common situation is to vary the assumed sample structure and recompute minimum detectable effects for two-level and three-level analyses. Tables 6 and 7 show what the implications would be for planning a study when the underlying cluster structure has twice as many classrooms per school as the study being used to compute the MDES (Table 6) and what the implications would be if the study being planned had half as many classrooms as the study being used to compute the MDES (Table 7). Note in all cases that the total number of schools as well as the total number of students per school remain constant, and thus the two-level estimates used to create Tables 6 and 7 are the same as those used to create Table 5.

Recall that the original SBPP Stanford math data had approximately four students per classroom and two classrooms per school (see Table 1). Table 6 explores the implications of planning a study in which, instead of having two classrooms per school and four students per classroom, there are instead four classrooms per school with two students per classroom. As before, the first set of columns in Table 6 show results for analyses without covariates. The second set of columns show analyses in which a school-level pretest is included and the third shows the results of analyses in which a student-level pretest is included.

The findings in Table 6 illustrate that when the number of classrooms per school is doubled and the number of students per school is held constant, the minimum detectable effect sizes computed from a two-level analysis with or without covariates are almost identical to those computed from a three-level analysis with the same data and covariates. Thus, if you are planning a study in which the number of classrooms per school is greater than the number in the study used to compute the MDES, using a two-

Table 6. Minimum Detectable Effect Sizes (MDES)
for Alternative Sample Structures, Double the Number of Classes per School

	Minimum Detectable Effect Size, Double the Number of Classes per School (Keep # of students/school constant, K=2Ko, N=No/2, # of schools = 60, T/C = 1:1)					
	No Covariates		School-Level Pretest		Student-Level Pretest	
Outcome	3-Level Model	Difference (3 lvl-2 lvl)	3-Level Model	Difference (3 lvl-2 lvl)	3-Level Model	Difference (3 lvl-2 lvl)
SBPP:						
Stanford 9 Total Math scaled score	0.336	-0.006	0.254	0.001	0.309	-0.008
Stanford 9 Total Reading scaled score	0.325	-0.015	0.250	0.002	0.319	-0.014
RFIS:						
SAT 10 reading comprehension test grade 1	0.247	-0.014	0.183	-0.018	0.203	-0.018
SAT 10 reading comprehension test grade 2	0.214	-0.010	0.148	-0.001	0.169	-0.016
SAT 10 reading comprehension test grade 3	0.211	-0.014	0.151	-0.015	0.173	-0.018
FL Elementary School Data:						
FCAT Math test for grade 5	0.263	-0.017	0.156	-0.022	0.149	-0.006
FCAT Reading test for grade 5	0.242	-0.015	0.140	-0.019	0.132	-0.004
NC Elementary School Data:						
Math test for grade 5	0.256	-0.016	0.193	-0.018	0.184	-0.008
Reading test for grade 5	0.232	-0.011	0.174	-0.012	0.154	-0.006
NC Secondary School Data:						
High School Algebra 2	0.322	-0.034	0.265	-0.032	0.224	-0.010
High School Biology	0.274	-0.035	0.217	-0.046	0.213	-0.019
High School Chemistry	0.272	-0.031	0.233	-0.025	0.208	-0.024
High School Geometry	0.350	-0.045	0.220	-0.029	0.219	-0.009
High School Physics	0.362	-0.025	0.289	-0.031	0.248	0.001

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated values for the intraclass correlations were obtained from a three-level model and a two-level model of the outcome measure without covariates. A school-level pretest and a student-level pretest measure were used in the model to obtain the R-squared values used in the MDES calculation for models with covariates. In addition, all analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

level model for analysis purposes will provide good estimates of the MDES, even though the middle level is not being accounted for explicitly.

Table 7 shows corresponding findings after halving the number of classrooms per school but holding constant the number of students per school. The results shown in Table 7 also indicate that, with the exception of the North Carolina secondary school data, the MDESs from the two- and three-level analyses yield quite comparable results, even though the sample structure has changed substantially. For the elementary school data, the difference between the estimates of MDES derived from the two- and three-level model are never more than 0.031. However, for the North Carolina secondary school data the differences between the estimates obtained from the two- and three-level analyses are much more sizable, ranging from 0.073 to 0.099. When a school-level pretest is added to the model (second set of columns) — a step that, as seen earlier, tends to magnify the difference between the two- and three-level models — the differences in MDES between the two models range from 0.093 to .0.126 for the various North Carolina secondary school outcomes. In this instance, using a two-level model to estimate the MDES in a study where the underlying data structure is actually comprised of three levels could be misleading. Yet, as also demonstrated earlier, including a student-level pretest (third set of columns) can reduce the difference between the estimates obtained from the two- and three-level models and help mitigate problems. In this case, the inclusion of the student-level pretest does reduce the differences substantially.

Planning a Study: Summary

Given these findings, what are the implications of planning a study that randomizes groups comprised of three levels of variation without explicitly accounting for the middle level? The preceding discussion shows that in almost all instances the MDES obtained using two levels of data (for example. students clustered within schools) is very similar to what would have been obtained with data at three levels (for example, students clustered within classrooms within schools). This is true even when the data being used for planning purposes have a variable cluster structure, include covariates at the student level or school level, or do not reflect the same underlying structure as the sample used in the actual study (that is, same number of students per classroom and classrooms per school). The similarity of MDESs is especially true for data in which the variance component at the classroom level is relatively small — which is usually the case in elementary schools. When the classroom-level variance component is large, the difference between the estimates derived from the two- and three-level analyses can in rare cases be meaningful, and the addition of a school-level pretest variable can make this problem worse. But including a pretest variable at the student level can help eliminate this problem under most circumstances.

Table 7. Minimum Detectable Effect Sizes (MDES) for Alternative Sample Structures, Half of the Number of Classes per School

	Minimum Detectable Effect Size, Half of the Number of Classes per School (Keep # of students/school constant, K=Ko/2, N=2No, # of schools = 60, T/C = 1:1)					
	No covariates		School-Level Pretest		Student-Level Pretest	
Outcome	3-level Model	Difference (3 lvl-2 lvl)	3-level Model	Difference (3 lvl-2 lvl)	3-level Model	Difference (3 lvl-2 lvl)
SBPP:						
Stanford 9 Total Math scaled score	0.352	0.011	0.260	0.007	0.329	0.012
Stanford 9 Total Reading scaled score	0.365	0.026	0.260	0.011	0.358	0.025
RFIS:						
SAT 10 reading comprehension test grade 1	0.278	0.017	0.224	0.023	0.241	0.020
SAT 10 reading comprehension test grade 2	0.251	0.026	0.192	0.042	0.215	0.030
SAT 10 reading comprehension test grade 3	0.253	0.028	0.204	0.038	0.224	0.032
FL Elementary School Data:						
FCAT Math test for grade 5	0.311	0.031	0.227	0.049	0.166	0.011
FCAT Reading test for grade 5	0.287	0.031	0.208	0.049	0.144	0.008
NC Elementary School Data:						
Math test for grade 5	0.303	0.031	0.251	0.040	0.203	0.011
Reading test for grade 5	0.267	0.024	0.218	0.032	0.164	0.006
NC Secondary School Data:						
High School Algebra 2	0.447	0.091	0.406	0.109	0.287	0.054
High School Biology	0.393	0.084	0.355	0.093	0.264	0.032
High School Chemistry	0.399	0.095	0.371	0.113	0.272	0.041
High School Geometry	0.469	0.073	0.374	0.126	0.268	0.040
High School Physics	0.486	0.099	0.434	0.114	0.317	0.070

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated values for the intraclass correlations were obtained from a three-level model and a two-level model of the outcome measure without covariates. A school-level pretest and a student-level pretest measure were used in the model to obtain the R-squared values used in the MDES calculation for models with covariates. In addition, all analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

Analyzing Data with a Three-Level Structure Using a Two-Level Model

Until now the discussion has focused on *planning* future studies using three-level data when the extant data lack information at the middle level, that is, the classroom level. We now consider the *analysis* of the data from the impact study itself: Specifically, does the point estimate and estimated standard error for an impact at the school level remain the same whether or not the middle level of a three-level situation is considered explicitly? This question is particularly important, since in many instances researchers are not able to explicitly link students to classes within schools and have no choice but to estimate a two-level model that does not explicitly consider the middle level of the data structure.

It has been shown that estimating a three-level model using feasible generalized least squares that fully account for the clustering in one's data will provide consistent and asymptotically efficient estimates (Cheung, Fotiu, and Raudenbush, 2001). The questions here are whether researchers can obtain consistent estimates of program impact if they misspecify the model by not explicitly accounting for the middle level of clustering and whether the resulting estimates will be asymptotically efficient.

It can be shown that for samples with a constant number of students per classroom and classrooms per school and no covariates at the student, classroom, or school level other than the treatment indicator at school level, one will obtain identical estimates of program impacts and identical estimates of standard errors whether or not the middle level of a three-level situation is explicitly acknowledged.[2] However, as was the case when MDESs obtained from two- and three-level models were considered, these proofs only hold for data that have a cluster structure that remains constant across clusters (that is, schools), which is rarely the case in practice. In addition, the proofs do not take into account situations in which covariates are included at the student or school level — a situation that also frequently occurs. Furthermore, the proofs are for expected values of the estimators being considered, not for specific estimates from a given sample. To explore how well conclusions from the proofs hold for a broader and more realistic range of data structures, the paper returns to its empirical analyses.

Table 8 shows coefficient estimates and estimated standard errors for a school-level treatment indicator using both a two- and three-level model for the four data sources in the study. As discussed in the data section, since specific interventions for the Florida and North Carolina data are unavailable, school-treatment status was randomly assigned so that half of the schools in those two states are in the treated

[2]Proof of this statement is available from the authors upon request.

Table 8. Three-Level vs. Two-Level Model Comparisons: Impact Estimates and Standard Errors (S.E.)

	Impact Estimate *without Covariates*				*with School Covariate*				*with Student Covariate*			
	3-Level Model		2-Level Model		3-Level Model		2-Level Model		3-Level Model		2-Level Model	
Outcome	**Impact**	**S.E.**	**Impact**	**S.E.**	**Impact**	**S.E.**	**Impact**	**S.E.**	**Impact**	**S.E.**	**Impact**	**S.E.**
SBPP:												
Stanford 9 Total Math scaled score	4.272	3.739	4.271	3.738	6.368	3.476	6.385	3.481	-1.511	2.688	-1.531	2.674
Stanford 9 Total Reading scaled score	3.299	4.303	3.433	4.300	4.538	4.270	4.699	4.256	-1.566	2.528	-1.537	2.542
RFIS:												
SAT 10 reading comprehension test grade 1	-0.099	2.350	-0.132	2.385	4.165	1.986	3.952	2.039	1.624	1.804	1.660	1.835
SAT 10 reading comprehension test grade 2	-3.348	1.728	-2.983	1.715	0.152	1.402	0.345	1.414	-0.051	1.064	0.108	1.072
SAT 10 reading comprehension test grade 3	-4.038	1.590	-3.749	1.594	-0.911	1.355	-0.750	1.375	0.057	0.795	0.134	0.799
FL Elementary School Data:												
FCAT Math test for grade 5	-0.012	0.019	-0.011	0.019	0.000	0.011	-0.007	0.012	0.005	0.010	-0.002	0.011
FCAT Reading test for grade 5	0.001	0.018	-0.001	0.018	-0.001	0.011	-0.011	0.012	0.006	0.010	-0.002	0.010
NC Elementary School Data:												
FCAT Math test for grade 10	-0.036	0.025	-0.031	0.027	-0.010	0.015	-0.006	0.014	-0.010	0.013	-0.017	0.011
FCAT Reading test for grade 10	-0.024	0.029	-0.021	0.031	-0.017	0.016	-0.016	0.015	-0.011	0.013	-0.019	0.011
NC Secondary School Data:												
High School Algebra 2	-0.055	0.072	-0.071	0.072	-0.074	0.061	-0.080	0.060	-0.049	0.049	-0.047	0.047
High School Biology	0.105	0.057	0.100	0.058	0.057	0.048	0.055	0.049	0.058	0.042	0.049	0.043
High School Chemistry	-0.089	0.082	-0.073	0.081	-0.092	0.072	-0.084	0.069	-0.039	0.060	-0.033	0.062
High School Geometry	-0.011	0.075	-0.015	0.077	0.027	0.051	0.022	0.048	-0.003	0.045	-0.007	0.044
High School Physics	0.202	0.258	0.140	0.249	0.165	0.217	0.108	0.206	0.166	0.173	0.146	0.157

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated impacts were obtained from a three-level model and a two-level model of the outcome measure with or without school or student-level pretests as covariate. All analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

group and the other half in the control group. The first four columns include no covariates other than a treatment-status indicator and indicators for districts. Columns 5 through 8 show models with a school-level pretest included and columns 9 through 12 include a student-level covariate.

While the point estimates and standard errors shown in Table 8 are not exactly the same for the two types of analyses, they are in most instances quite comparable whether or not covariates are included at either the school or student level. Even in the instances where the point estimates and standard errors differ somewhat, the same inferences would be drawn from a two-level or a three-level model. For example, for the RFIS data the impact estimate for the second-grade test was -3.35 with a standard error of 1.73 when the impact was estimated using a three-level model. The corresponding two-level model yielded an impact estimate of -2.98 with a standard error of 1.72. Both point estimates are similar in magnitude and neither is statistically significant, so in both instances the evidence indicates that Reading First had no impact on second-grade SAT 10 reading scores. These findings hold across data sets of widely varying sizes and structures.

As was the case for planning a study, these findings suggest that a two-level model can be used to estimate program impacts even when it does not explicitly acknowledge a middle level of clustering. This is particularly true when the middle-level variance component is small, as is the case for most elementary school outcomes. However, the finding also holds for secondary school data, where the classroom-level variance component is relatively larger, and for situations where the cluster structure varies across the schools in the sample and when covariates are included in the model.

Conclusions

As noted, this paper is intended to provide practical guidance to researchers who are designing and analyzing studies that randomize schools to measure intervention effects on student academic outcomes when no information is available about the middle (classroom) level of clustering. Using four multisite data sets based on academic outcomes for students within classrooms within schools, the paper has explored in detail the implications of not explicitly acknowledging the middle level when planning or analyzing data in which the coefficient of interest is at the third (school) level. The analysis shows that in almost all situations one will obtain nearly identical results whether or not the classroom or middle level is acknowledged explicitly. With one exception, this conclusion holds for both elementary school data (for which the classroom variance component is typically quite small) and for secondary school data (for which the classroom variance component is somewhat larger), for data sets with varying numbers of

student per classroom and classrooms per school, in situations where covariates are included at either the student or school level, and in situations where the cluster structure of the study being planned differs substantially from the one used for planning purposes. The only potential problem arises when the middle-level variance component is large (which is usually only the case for secondary school data) and when the study being planned has a markedly different cluster structure than the study that was used for planning purposes. Even in this kind of situation, if a student-level pretest variable is included in the models, any potential problems that may arise can be virtually eliminated. Thus in most situations researchers can proceed with two-level analyses of three-level data without too much cause for concern.

Appendix A

Three-Level vs. Two-Level Model Comparisons: Estimated R-Squared for School- and Student-Level Covariates

Appendix Table A-1. Three-Level vs. Two-Level Model Comparisons: Estimated R-Squared for School- and Student-Level Covariates

	Estimated R-Squared for School-Level Covariate					Estimated R-Squared for Student-Level Covariate				
	3-Level Model			2-Level Model		3-Level Model			2-Level Model	
Outcome	School-Level	Class-Level	Student-Level	School-Level	Student-Level	School-Level	Class-Level	Student-Level	School-Level	Student-Level
SBPP:										
Stanford 9 Total Math scaled score	0.377	-0.107	0.005	0.311	0.004	0.514	0.393	0.477	0.510	0.475
Stanford 9 Total Reading scaled score	0.092	0.057	-0.003	0.091	-0.001	0.765	0.890	0.503	0.795	0.520
RFIS:										
SAT 10 reading comprehension test grade 1	0.506	-0.021	0.000	0.379	0.000	0.504	0.351	0.139	0.470	0.150
SAT 10 reading comprehension test grade 2	0.752	-0.022	0.000	0.503	0.000	0.659	0.711	0.464	0.661	0.477
SAT 10 reading comprehension test grade 3	0.674	-0.006	0.000	0.420	0.000	0.846	0.798	0.522	0.830	0.536
FL Elementary School Data:										
FCAT Math test for grade 5	0.824	0.009	0.000	0.649	0.000	0.663	0.805	0.605	0.696	0.628
FCAT Reading test for grade 5	0.874	0.014	0.000	0.687	0.000	0.697	0.868	0.505	0.738	0.543
NC Elementary School Data:										
Math test for grade 5	0.592	0.012	0.000	0.460	0.000	0.406	0.718	0.654	0.475	0.659
Reading test for grade 5	0.615	0.015	0.000	0.499	0.000	0.515	0.827	0.566	0.580	0.579
NC Secondary School Data:										
High School Algebra 2	0.493	0.014	0.000	0.322	0.000	0.461	0.661	0.329	0.576	0.467
High School Biology	0.675	0.003	0.000	0.312	0.000	0.229	0.693	0.310	0.442	0.422
High School Chemistry	0.484	0.023	0.000	0.305	0.000	0.278	0.636	0.308	0.419	0.403
High School Geometry	0.846	0.057	0.000	0.633	0.000	0.563	0.755	0.391	0.674	0.526
High School Physics	0.526	0.008	0.000	0.340	0.000	0.504	0.630	0.329	0.603	0.438

SOURCES: The School Breakfast Pilot Project (SBPP) first follow-up year database, the Reading First Impact Study (RFIS) first follow-up year database, the Florida Department of Education's K-20 Education Data Warehouse (FL-EDW) for 2005, and the North Carolina Education Research Data Center for 2005. Classes with only one student in the sample and schools with only one class in the sample are excluded from the analysis presented in this table.

NOTES: Estimated values for the R-squared were obtained from a three-level model and a two-level model of the outcome measure without covariates. All analyses include an indicator variable distinguishing treatment and control groups as well as indicator variables for school districts in the study sample.

References

Abt Associates Inc., and Promar International. 2005. *Evaluation of the School Breakfast Program Pilot Project: Final Report.* Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service, Office of Analysis, Nutrition, and Evaluation.

Bloom, H. S. 2005. "Randomizing Groups to Evaluate Place-Based Programs." Pages 115-172 in Howard S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, H. S., L. Richburg-Hayes, and A. Black. 2007. "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis* 29: 30-59.

Cheong, Y. F., R. P. Fotiu, and S. W. Raudenbush. 2001. "Efficiency and Robustness of Alternative Estimators for Two- and Three-Level Models: The Case of NAEP." *Journal of Educational and Behavioral Statistics* 26 (4): 411- 429.

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

Hedges, L. V., and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29 (1): 60-87.

Moerbeek, M. 2004. "The Consequence of Ignoring a Level of Nesting in Multilevel Analysis." *Multivariate Behavioral Research* 39: 129-149.

Opdenakker, M. C., and J. Van Damme. 2000. "The Importance of Identifying Levels in Multilevel Analysis: An Illustration of the Effects of Ignoring the Top or Intermediate Levels in School Effectiveness Research." *School Effectiveness and School Improvement* 11: 103-130.

Tranmer, M., and D. G. Steele. 2001. "Ignoring a Level in a Multilevel Model: Evidence from UK Census Data." *Environment and Planning A* 33 (5): 941-948.

Van Landeghem, G., B. De Fraine, and J. Van Damme. 2005. "The Consequence of Ignoring a Level of Nesting in Multilevel Analysis: A Comment." *Multivariate Behavioral Research* 40: 423-434.

Van den Noortgate, W., M. C. Opdenakker, and P. Onghena. 2005. "The Effects of Ignoring a Level in Multilevel Analysis." *School Effectiveness and School Improvement* 16: 281-303.

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

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.