MDRC Working Papers on Research Methodology

Empirical Issues in the Design of Group-Randomized Studies to Measure the Effects of Interventions for Children

Howard Bloom[a]
Pei Zhu[a]
Robin Jacob[c]
Stephen Raudenbush[b]
Andres Martinez[c]
Fen Lin[b]

[a]**MDRC**
[b]**University of Chicago**
[c]**University of Michigan**

July 2008

Acknowledgments

This working paper is part of a series of publications by MDRC on alternative methods of evaluating the implementation and impacts of social and educational programs and policies. The authors thank Charles Michalopoulos and Alison Rebeck Black for their helpful comments. The working paper was supported by funding from the W. 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, Rockefeller, Jerry Lee, Spencer, William T. Grant, and Grable Foundations. Findings and conclusions in the paper do not necessarily represent the positions or policies of the funders.

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, Bristol-Myers Squibb Foundation, The Kresge Foundation, and The Starr Foundation. MDRC's dissemination of its education-related work is supported by the Bill & Melinda Gates Foundation, Carnegie Corporation of New York, and Citi 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, The Sandler Family Supporting Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report 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 studies that randomize groups to measure the impacts of interventions on children. To do so, the paper: (1) provides new empirical information about the values of parameters that influence the precision of impact estimates (intra-class correlations and R-squares); (2) examines the implications of planning group-randomized studies for three-level hierarchical situations, using empirical information obtained by estimating two-level hierarchical models (which under many conditions appears to not be problematic); and (3) assesses the implications of the uncertainty that exists when the design of group-randomized studies is based on estimates of intra-class correlations. Data for the paper come from two studies: the Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs) and the School Breakfast Pilot Project (SBPP). The analysis sample from CLIMBERs comprised 430 4-year old children from 47 preschool classrooms in 23 Chicago public schools. The analysis sample from the SBPP study comprised 1,151 third-graders from 233 classrooms in 111 schools in six school districts.

Contents

List of Tables and Figures

Table

Figure

Part I
Introduction

This paper addresses several empirical issues in the design of group-randomized studies to measure the effects of interventions on outcomes for children. Group-randomized studies have recently become a popular way to measure the effects of interventions (see Boruch and Foley, 2000, for a review and Boruch, 2005, for detailed examples). This approach randomizes intact groups, such as communities, hospitals, firms, schools, or child care centers, to treatment and control groups, data for which can provide unbiased estimates of intervention effects. In this way, randomizing groups is similar to randomizing individuals.

Most early applications of group randomization were in the health sciences. Consequently, the two existing textbooks on the approach focus on health research (Donner and Klar, 2000; Murray, 1998). Recently, several papers have attempted to make the approach accessible to other researchers (for example, Bloom, 2005; Schochet, 2005; Murray and Blitstein, 2003; Raudenbush, 1997). In addition, the current emphasis on randomized trials in education stimulated by the U.S. Department of Education's Institute of Education Sciences has prompted a series of large-scale group-randomized studies (for example, American Institutes for Research and MDRC, 2006). Correspondingly, a recent issue of *Education Evaluation and Policy Analysis* is comprised entirely of articles on the design of group-randomized studies (Raudenbush, Martinez, and Spybrook, 2007; Bloom, Richburg-Hayes, and Black, 2007; Hedges and Hedberg, 2007).

The core design decisions for group-randomized studies involve choosing: (1) the total number of groups to randomize; (2) the average number of individuals per group to observe; (3) the proportion of groups to allocate to treatment or control status; (4) what variables, if any, to use for covariate adjustments; and (5) what categories, if any, by which to block groups before they are randomized. Further design decisions are required for any given study, based on its specific goals and context.

Previous authors (for example, Raudenbush, 1997; Murray, 1998; Donner and Klar, 2000; Bloom, 2005) have presented the statistical framework of group-randomized studies. Group-randomization has a multilevel variance and covariance structure with individual subjects (level 1) clustered in randomized groups (level 2). For example, studies that measure impacts on students (level 1) by randomizing schools (level 2) have at least a two-level variance structure. The level 1 variance represents how an outcome varies across individual subjects (students) within randomized groups

(schools). The level 2 variance represents how the mean value of the individual outcome varies across randomized groups. The level 1 variance is often designated as σ^2, and the level 2 variance is often designated as τ^2. Using this framework, the total individual variance across all subjects in all randomized groups equals $\tau^2+\sigma^2$.

To design group-randomized studies that can attain desired levels of precision requires information about the variance at each level of a given situation. A simple two-level structure with individuals clustered in randomized groups requires knowledge of two variances, σ^2 and τ^2. This information is often expressed as the relationship between the two variances, referred to as an intra-class correlation, ρ, (ICC, Fisher, 1925), where:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \qquad (1)$$

The intra-class correlation is thus the proportion of total individual subject-level variance that is *between* randomized groups.

In addition, because it is often possible to markedly increase the precision of group-randomized studies by adjusting for baseline covariates (for example, Bloom, Richburg-Hayes and Black, 2007; Murray and Blitstein, 2004), knowledge of the predictive power of such covariates is essential for designing these studies. The predictive power of a covariate represents the proportion of the variance component at each level that is predicted (or "explained") by the covariate. These parameters are often referred to as R-squared values.

The statistical framework for group-randomized studies indicates how the core design decisions outlined above, together with the variance and covariance structure of the data to be analyzed, determine the statistical precision or power of impact estimates. The variance and covariance structure of the data depends on the type of group to be randomized (schools, communities, or hospitals, etc.) and the specific outcome measure or measures to be used (student achievement, individual behavior, or health status, etc.). It is therefore an empirical question as to how much precision a particular design will yield when used to address a specific impact question for a given target group. Consequently, a sound empirical foundation is needed to support the future development of fields that will use group-randomized studies. This foundation requires information about the variance and covariance structure of outcome measures for key target populations and types of randomized groups.

Information about the values of intra-class correlations and, to a lesser extent, R-squared values, has been catalogued by researchers in the health and prevention

sciences (for example, Murray and Short, 1995; Murray and Blitstein 2003; Siddiqui, Hedeker, Flay, and Hu, 1996). A repository of this information is maintained by David Murray and his associates.[1] Much less information is available for studies in education and child development, and most of this information is limited to outcome measures based on standardized achievement test scores (Bloom, Richburg-Hayes, and Black, 2007; Schochet, 2005; Hedges and Hedberg, 2007). Furthermore, most of the existing information is based on two-level data for students clustered in schools. This ignores the clustering of students in classrooms.

This paper attempts to expand the empirical foundation for designing group-randomized studies in education and child development, using two data sets derived from group-randomized studies. Part II describes the two data sources and defines the outcome measures examined. Part III presents estimates of intra-class correlations and R-squared values for a series of academic and child outcome measures. It also provides information for the three-level variance structure of these outcome measures. This three-level variance structure represents the clustering of students within classrooms and classrooms within schools. Part IV examines the statistical implications of the often necessary practice of designing studies that randomize schools as if they represent variances at two levels (students and schools), instead of three levels (students, classrooms, and schools). The question addressed here is: By how much has the design of these studies been misguided by this simplification? Fortunately, the surprising answer to this question is that, in most cases, the designs have probably not been misguided. Lastly, Part V examines the amount of uncertainty that exists for estimates of intra-class correlations from samples of different sizes and structures and explores the implications of this uncertainty for projections of the statistical precision of research designs.

[1]See http://sph.osu.edu/divisions/epidemiology/epifacstaff/murrayd/group-randomized-trials/

Part II
Data Sources, Student Samples, and Outcome Measures

Data for this paper were obtained from two studies that randomized schools to measure intervention effects on children: (1) the Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs)[2] and (2) the School Breakfast Pilot Project (SBPP) study.[3] This section describes the two studies, their samples, and their outcome measures used for the present paper.

Studies and Samples

CLIMBERs: This five-year study (2004-2009) is an evaluation of Breakthrough to Literacy, an early literacy curriculum, taken to scale in Chicago Public Schools preschool classrooms that serve 4-year-old children. Schools were recruited to participate in the study if they were low performing and had few other early literacy initiatives. Forty-four schools agreed to participate and were randomly assigned to a treatment group that implemented Breakthrough to Literacy or to a control group that did not implement the program. The goal of the project was to measure the impact of Breakthrough to Literacy at scale on students' preliteracy skills.

Participating schools served a largely low-income population; on average, 88 percent of their students came from low-income families. The schools also primarily served students of color; 86 percent reported that more than half their students were either African-American or Hispanic. Schools were typically large, with an average enrollment of 774 students and a range of 139 to 1,969. Annual mobility rates were high, averaging 23 percent and ranging from 7 percent to 56 percent.

One control school dropped out of the study prior to baseline data collection, and one treatment school dropped out prior to follow-up data collection. Because a central goal of this paper is to examine three-level data structures for students clustered within classrooms clustered within schools, only schools with two or more classrooms in the study are included. This limited the present analysis sample to 430 preschool stu-

[2]Abt Associates Inc., along with its research partners at the University of Iowa, is conducting the study, which is supported by a grant from the Institute for Education Sciences at the U.S. Department of Education.

[3]Abt Associates Inc., along with its research partner, Promar International, conducted the study under contract to the Office of Analysis, Nutrition, and Evaluation at the U.S. Department of Agriculture, Food and Nutrition Service.

dents from 47 classrooms in 23 schools. Data for these students were used to estimate intra-class correlations for three-level and two-level variance structures and corresponding R-squared values based on student scores from a standardized preliteracy test.

SBPP:[4] This three-year demonstration project (2000-2003) was based on an experimental design that randomized schools within six school districts (in Alabama, Arizona, California, Idaho, Kansas, and Mississippi) to a treatment condition, in which schools implemented a universal free school breakfast program, or to a control condition, in which schools continued to operate their regular school breakfast programs for eligible children from low-income families.[5] The goal of the project was to measure the added value of universal free school breakfasts.

The six school districts were chosen for the project from among 136 that applied. The resulting project sample included students in grades 2 through 6 from 138 elementary schools. Within each treatment school or control school, six classrooms were selected randomly for analysis, with at least one classroom per grade. This paper uses data for third-grade students, because the sample for this grade is by far the largest and most complete. The findings reported are based on data for 1,151 third-graders from 233 classrooms in 111 schools located in 6 school districts.[6] The outcomes measures were obtained from several sources and focused on academic outcomes, other school-related outcomes, emotional and behavioral outcomes, as well as health outcomes. These data were used to estimate intra-class correlations for three-level and two-level variance structures and corresponding R-squared values from the use of covariates (described later).

Measures

Outcome measures for this paper comprise four categories: (1) academic outcomes (standardized test scores); (2) other school-related outcomes (for example, attendance); (3) student behavior; and (4) health outcomes. These measures are briefly described below. For further details, see Appendix A.

Academic Outcomes: Four measures of preliteracy skills were obtained from data for CLIMBERs, based on student scores from the Preschool Comprehensive Test

[4]The discussion in this part is based on Abt Associates Inc. and Promar (2005).

[5]The pilot program used a matched-pair random assignment design with schools as the unit of random assignment.

[6]The number of students, classrooms, and schools vary by outcomes due to item nonresponse.

of Phonological and Print Processing (Lonigan, Wagner, Torgesen, and Rashotte, 2002)[7]:

- "Print Awareness" measures beginning knowledge about written language, for example, knowing what print looks like and how it works.
- "Elision" tests a child's ability to segment spoken words into smaller parts by deleting parts and then recalling the remaining portion.
- "Blending" measures a child's ability to put sounds together to form words. For example, "What word do these sounds make: 't-oi' "?
- "Expressive Vocabulary" measures the number of different words a child uses when speaking or writing.

Two measures of third-grade academic performance were obtained from data for the SBPP. These measures come from the Stanford Achievement Test Series, Ninth Edition (SAT 9):

- Total scaled score for mathematics
- Total scaled score for reading

Other Academic-Related Outcomes: The SBPP also collected supplementary measures of student academic performance. Measures used in this paper include:

- Attendance and tardiness
- Participation in school breakfasts
- "Stimulus Discrimination" (Detterman, 1988), which comprises three measures of cognitive performance
- The "Digit Span" subtest of the Wechsler Intelligence Scales for Children III (Wechsler, 1991), which assesses short-term auditory memory
- Tasks of "verbal fluency" that count the number of items that students name in a given period of time to test their longer-term memory (Simeon and Grantham-McGregor, 1989)

[7]This test measures phonological skills that have been shown to be important precursors to reading proficiency. The test has not yet been published, and there is little information about its psychometric properties — but it is used widely with middle-income and low-income students.

Emotional and Behavioral Outcomes: The SBPP provides a series of psychosocial and behavioral measures for young children. Measures used for this paper are:

- "Social and emotional functioning," assessed through the Pediatric Symptom Checklist (PSC) (Murphy et al., 1998), administered as part of a survey of students' parents
- "Behavioral measures" from the Conners' Teacher Rating Scales-Revised CTRS-R(s),[8] which consists of 28 questions on which teachers rate their students. These questions are used to create four scales: Attention Deficit Hyperactivity Disorder (ADHD) Index, Cognitive Problems/Inattention, Hyperactivity, and Oppositional Behavior.
- "Behavioral measures" from the Children's Behavior Questionnaire, which measures children's temperaments (Rothbart, Ahadi, and Evans, 2000). Two subscales, Ability to Focus and Ability to Follow Instructions, are used.

Health Outcomes: In addition, the SBPP collected a series of measures of student's health status. Measures used for this paper include:

- The Body Mass Index
- Indicators for weight status, whether a child is considered "overweight" or "at risk of overweight"
- Height
- Weight

[8]The CTRS-R(s) is a part of a larger set of measures, the Conners' Rating Scales, which have long been used to assess psychopathology and behavior issues, such as problems with conduct, anxiety, and social functioning, as well as Attention Deficit Hyperactivity Disorder (ADHD) in children and adolescents (Conners, 2000).

Part III
New Information about
Intra-Class Correlations and R-Squared Values

This part of the paper describes how data on intra-class correlations and R-squared values can be used to design group-randomized studies. It also presents estimates of these parameters from data for the two studies described above and illustrates the implications of these estimates for the statistical precision of alternative sample designs.

Precision of Impact Estimates

One of the most important features of an impact study is its ability to provide adequate precision for estimates of intervention effects. This paper reports precision as a minimum detectable effect size (MDES), which, intuitively, is the smallest true intervention effect that a study sample can detect with confidence. Conventionally, a MDES is defined as the smallest true program impact that would have an 80 percent chance of being detected (80 percent statistical power) with a two-tailed hypothesis test at the 0.05 level of statistical significance. This paper follows this convention.

To choose a MDES for a given study requires an understanding of its specific circumstances. For example, from a benefit-cost perspective one might ask whether a proposed sample could reliably detect the smallest impact required for an intervention to "break even" (that is, produce benefits equal to its costs). In other words, one would want a sample that was large enough to ensure that an estimated impact near the "break-even" point would be reliable. A smaller sample could detect only much larger impacts, which might be impossible to attain. Hence, this smaller sample would be "underpowered" statistically. Hill, Bloom, Black, and Lipsey (forthcoming) provide a series of empirical benchmarks for helping to determine an appropriate MDES for educational interventions. There is little such empirical guidance for other fields of intervention research, however.

A MDES is defined in terms of the underlying population's standard deviation for a given outcome measure. For example, a MDES of 0.20 for student achievement indicates that an impact analysis can reliably detect a program-induced increase in student achievement that is equal to or greater than 0.20 standard deviation of the existing student outcome distribution. Mathematically, a MDES is proportional to the standard

error of the impact estimate and to the inverse of the underlying population's standard deviation for the outcome. This relationship can be expressed as:

$$MDES = M * \sqrt{Var(impact)} / \sigma_{total} \quad (1)$$

Where M is a multiplier that depends on the assumed power, significance level, and one- or two-tail nature of the statistical test, plus the number of degrees of freedom of the study design,[9] Var (impact) is the variance of the impact estimate and σ_{total} is the standard deviation of the outcome measure across all individual subjects in the target population (or sample).

For group-randomized designs, the standard errors of impact estimates are larger (often by a lot) than those for individual-randomized designs for the same total number of individuals (Bloom, 2005). This is because the clustering of students within classrooms and schools causes differences in average outcomes across schools (the school-level variance component) and/or classrooms (the classroom-level variance component) to increase the standard error of impact estimates under group randomization by more than under individual randomization.

Consequently, variance expressions for a group-randomized design must account for each variance component. For example, the MDES for a study that randomizes schools and has a three-level data structure with students clustered within classrooms and classrooms clustered within schools is as follows, assuming no covariates:

$$MDES = \frac{M_{(J-2)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau^2}{J} + \frac{\gamma^2}{J*K} + \frac{\sigma^2}{J*K*N}} * \frac{1}{\sqrt{\tau^2 + \gamma^2 + \sigma^2}} \quad (2)$$

where $M_{(J-2)}$ = a multiplier defined in Appendix B;

P = the proportion of schools assigned to the treatment group;

τ^2 = the unconditional variance (without covariates) of mean outcomes across schools;

γ^2 = the unconditional variance (without covariates) of classroom means within schools;

σ^2 = the unconditional variance (without covariates) of student outcomes within classrooms;

[9]For a two-group experimental design without covariates, the number of degrees of freedom equals the number of randomized groups minus the two parameters in the model or J-2. The magnitude of M decreases as J increases. See Appendix B for a detailed definition of M.

J = the total number of schools randomized to treatment or control status;
K = the harmonic mean number of classrooms per school; and
N = the harmonic mean number of students per classroom.

Equation 2 corresponds to Equation 1 in that:

- $(\tau^2 + \gamma^2 + \sigma^2)$ equals the total variance of the outcome measure across all students from all classrooms in all schools, or σ^2_{total}.
- $(\frac{\tau^2}{J} + \frac{\gamma^2}{J*K} + \frac{\sigma^2}{J*K*N}) * \frac{1}{P(1-P)}$ represents the influence of the school-level, classroom-level, and student-level variance components and the proportion of clusters randomized to treatment status.

In practice, baseline characteristics, such as students' prior test scores and demographics, are often used as covariates in a multilevel regression model to improve the precision of impact estimates. Such models (described later) estimate the intervention effect as a regression-adjusted difference of mean outcomes for the treatment and control groups. 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 MDES is:

$$MDES = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau^2(1-R^2_{sc})}{J} + \frac{\gamma^2(1-R^2_{cl})}{J*K} + \frac{\sigma^2(1-R^2_{st})}{J*K*N}} * \frac{1}{\sqrt{\tau^2 + \gamma^2 + \sigma^2}} \qquad (3)$$

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 before.

Here 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,

$$R_L^2 = \frac{\sigma_{U,L}^2 - \sigma_{C,L}^2}{\sigma_{U,L}^2} \quad (4)$$

where $\sigma_{U,L}^2$ is the unconditional variance at level L without covariates,

$\sigma_{C,L}^2$ is the conditional variance at level L when covariates are added.

Note that when there are no covariates, all R-squared values equal zero and Equation 3 reduces to Equation 2. On the other hand, by including covariates, unexplained variance can, in some cases, be reduced and precision can be improved. It is also possible that, under certain circumstances, the inclusion of covariates at level 1 can increase the unexplained variation at level 2 or level 3 and thereby decrease precision. This increase in unexplained variation at level 2 or level 3 would be reflected by a negative value for the relevant R-squared.

Relationships among τ^2, γ^2, and σ^2 can be expressed as intra-class correlations like that in Equation 1. The intra-class correlation at the school level ρ_{sc} equals the proportion of total student variance ($\tau^2 + \gamma^2 + \sigma^2$) that is between schools. The intra-class correlation at the classroom level, ρ_{cl}, equals the proportion of total student variance that is between classrooms within schools. In symbols:

$$\rho_{sc} = \frac{\tau^2}{\tau^2 + \gamma^2 + \sigma^2}$$

and

$$\rho_{cl} = \frac{\gamma^2}{\tau^2 + \gamma^2 + \sigma^2}$$

The remaining proportion of total student variance $(1 - \rho_{sc} - \rho_{cl})$ is the variance between students within a class. Therefore, an alternative way to express the MDES for a three-level variance structure is:

$$MDES = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\rho_{sc}(1-R_{sc}^2)}{J} + \frac{\rho_{cl}(1-R_{cl}^2)}{J*K} + \frac{(1-\rho_{sc}-\rho_{cl})(1-R_{st}^2)}{J*K*N}} \quad (5)$$

where all parameters are defined as before.

Equation 5 provides a simple way to assess the precision of alternative sample designs. But to do so requires information about the school-level and classroom-level

intra-class correlations and the school-level, classroom-level, and student-level R-squared values.

Estimation Model

This section describes how values for the preceding parameters were estimated from data for the Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs) and the School Breakfast Pilot Project (SBPP). Because data from both studies identify students within classrooms within schools, variance components and R-squared values were estimated using the following three-level hierarchical model:[10]

Level 1:

$$Y_{ijk} = \pi_{0jk} + \sum_{s>0} \pi_{sjk} X_{sijk} + \varepsilon_{ijk} \qquad (6)$$

where:

Y_{ijk} = the value of the outcome measure for student i from classroom j in school k;

π_{0jk} = the regression-adjusted mean value of the outcome measure for classroom j in school k;

X_{sijk} = the value of the s^{th} student-level covariate for student i from classroom j in school k; and

ε_{ijk} = the residual error for student i from classroom j in school k, which is assumed to be independently and identically distributed.

Level 2:

$$\pi_{0jk} = \beta_{0k} + \gamma_{jk} \qquad (7)$$

where:

β_{0k} = the mean value of the outcome measure for school k and

γ_{jk} = the residual error for classroom j from school k, which is assumed to be independently and identically distributed.

[10]All models were estimated by Restricted Maximum Likelihood Estimation, using the PROC MIXED procedure in SAS.

Level 3:

$$\beta_{0k} = \theta_0 + \theta_1 T_k + (\sum_{m>1} \theta_m Z_{mk}) + \mu_k \quad (8)$$

where:

θ_0 = the grand mean of the regression-adjusted outcome measure for the average control school;

T_k = one for treatment schools and zero for control schools;

θ_1 = the estimated impact of treatment;

Z_{mk} = the m^{th} school-level covariate for school k; and

μ_k = the residual error for school k, which is assumed to be independently and identically distributed.

By including an indicator variable for treatment or control status (Tk) this model removes all existing differences between the treatment and control groups (treatment effects) when estimating variance components. In addition, for the SBPP, the model removes all differences among the six participating school districts by including indicator variables for them as school-level covariates (Zmk). Hence, all estimates represent within-district variances in the absence of treatment effects.

The first step in the analysis for an outcome measure was to estimate the preceding model without covariates in order to estimate its unconditional variance components (τ^2, γ^2 and σ^2). The second step was to compute the school-level and classroom-level unconditional intra-class correlations (ρ_{sc} and ρ_{cl}) from the estimated unconditional variance components. The third step was to estimate values for each conditional variance component using a model that included covariates. The final step was to compute R-squared values for each level $(R^2_{sc}, R^2_{cl}, R^2_{st})$ by comparing the magnitudes of its conditional and unconditional variance components.

Key Findings

Table 1 lists parameter estimates for all outcome measures in the analysis. The first two columns list school-level and classroom-level unconditional intra-class correlations (estimated without covariates). As noted before, the remaining proportion of the total variance comes from variance between students within a class; the last three columns list school-level, classroom-level, and student-level R-squared values (obtained by comparing estimates of conditional and unconditional variance components). Findings for academic outcomes are from the CLIMBERs preschool sample and the SBPP third-grade sample. Findings for other outcomes are from the SBPP third-grade sample.

Table 1 Parameters Estimated from a Three-Level Model

Outcome	Unconditional ICC		R-squared		
	School	Class	School	Class	Student
Academic Outcomes					
Print Awareness[b] (CLIMBERs)	0.308	0.016	0.580	0.000	0.000
Blending[b](CLIMBERs)	0.149	0.011	0.346	0.000	0.000
Elision[b] (CLIMBERs)	0.000	0.068	n.e.	0.000	0.000
Expressive Vocabulary[b](CLIMBERs)	0.055	0.091	1.000	0.000	0.000
Stanford 9 total math scaled score[a,c]	0.081	0.026	0.494	0.627	0.482
Stanford 9 total reading scaled score[a,c]	0.059	0.086	0.840	0.880	0.510
Academic-Related Outcomes					
Breakfast participation (adjusted for attendance)[a,c]	0.206	0.000	0.385	n.e.	0.320
Attendance[a,c]	0.000	0.060	n.e.	0.525	0.311
Days tardy as a percentage of number of school days enrolled[c]	0.077	0.000	0.253	n.e.	0.217
Stimulus Discrimination: number of trials incorrect[c]	0.000	0.051	n.e.	-0.001	-0.002
Stimulus Discrimination: average trial time[c]	0.049	0.044	0.267	0.163	0.020
Stimulus Discrimination: average viewing time[c]	0.045	0.044	0.271	0.176	0.017
Digit Span: forward and backward combined and scaled by age[c]	0.022	0.000	0.258	n.e.	0.049
Verbal Fluency: number of animals named[c]	0.053	0.046	0.670	0.029	0.025
Verbal Fluency: number of things to eat named[c]	0.040	0.044	0.791	-0.132	0.025
Verbal Fluency: VF_ani and VF_eat combined[c]	0.054	0.046	0.771	-0.068	0.033
Emotional and Behavioral Outcomes					
Pediatric Symptom Checklist (PSC) status, 0=non-PSC case 1=PSC case[c]	0.000	0.000	-3.128	n.e.	0.021
Sum of 17 PSC questions[c]	0.021	0.021	-0.231	0.207	0.042
Conners' ADHD Index[c]	0.008	0.078	0.699	-0.054	0.038
Cognitive Problems/Inattention[c]	0.005	0.033	1.000	0.279	0.083
Hyperactivity[c]	0.000	0.074	n.e.	0.026	0.019
Oppositional Behavior[c]	0.000	0.037	n.e.	0.139	0.037
Ability to Focus[c]	0.001	0.125	1.000	-0.008	0.104
Ability to Follow Instructions[c]	0.000	0.130	n.e.	0.017	0.120
Health Outcomes					
Body Mass Index percentile[c]	0.000	0.000	n.e.	n.e.	0.004
At risk of overweight[c]	0.006	0.000	0.363	n.e.	0.002
Considered overweight[c]	0.000	0.035	n.e.	-0.029	0.002
Weight status[c]	0.003	0.007	0.231	0.014	0.003
Height[c]	0.017	0.008	1.000	-0.162	0.048
Weight[c]	0.017	0.018	0.574	-0.470	0.016

Sources: Where indicated, data are from the CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the intra-class correlations were obtained from a three-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for each school district in the study sample.

[a]Baseline measure of the outcome variable is included as prior achievement measure in the model.

[b]Baseline measure of other academic outcomes is included as prior achievement measure in the model.

[c]Student-level demographic information (age, ethnicity, gender, eligibility for free/reduced lunch) is included in the model.

n.e.=not estimable.

n.a.=not available.

Unconditional Intra-Class Correlations

For academic outcomes, the majority of school-level unconditional intra-class correlations range from about 0.06 to 0.15, and all classroom-level unconditional intra-class correlations are less than 0.10. The mean value of the unconditional intra-class correlation is 0.11 for schools and 0.05 for classrooms.

For three academic outcomes (Print Awareness, Blending, and the SAT 9 math test), the school-level intra-class correlation exceeds the classroom-level intra-class correlation. This may reflect the fact that schools in the sample serve different student populations. For three of the academic outcomes (Elision, Expressive Vocabulary, and the SAT 9 reading test), the classroom intra-class correlation is larger than the school intra-class correlation. This might reflect that fact that certain skills are influenced more by teacher characteristics than by school conditions.

Mean values for school-level and classroom-level unconditional intra-class correlations are 0.05 and 0.03, respectively, for academic-related outcomes, such as school breakfast program participation, school attendance, Stimulus Discrimination, Digit Span and Verbal Fluency. Of the 10 outcome measures in this category, three have estimated intra-class correlations that equal zero for classrooms, and two have estimated intra-class correlation that equal zero for schools. Values for the remaining measures at both the classroom level and school level are typically less than 0.05.

For emotional and behavioral outcome measures, the mean value of the unconditional intra-class correlation is less than 0.01 for schools and approximately 0.06 for classrooms. For all of these outcome measures, the classroom intra-class correlation is larger than that for schools. This is perhaps because the measures were constructed from teacher ratings.[11]

For health measures, the mean intra-class correlation is less than 0.01 for schools and approximately 0.01 for classrooms. These small magnitudes may reflect the fact that young students have had limited exposure to school environmental and contextual factors that could shape their physical development.

At this point it is useful to ask: How do the present findings compare with those from previous research? As noted, there are only a few studies that provide such information. Hedberg, Santana, and Hedges (2004) report unconditional school-level intra-class correlations for academic outcomes based on data for several large national samples. These values typically range from about 0.15 to 0.30 and reflect differences in

[11]However, the Pediatric Symptom Checklist questions were answered by parents.

outcomes that exist across both within and across school districts. Based on evidence from past empirical studies and new evidence from three evaluation studies, Schochet (2005) concludes that: "the examined data sources suggest that values for ρ_1 (*which we refer to as the unconditional school-level intra-class correlation within a district*) often range from 0.10 to 0.20 for standardized test scores." Bloom, Richburg-Hayes, and Black (2007) report school-level intra-class correlations that range from about 0.15 to 0.20 for reading and math test scores, using third-grade data from five urban school districts.

There is also a large and growing body of empirical research on the magnitudes of intra-class correlations for public health outcomes and the incidence of risk behaviors — such as smoking, drinking, drug abuse, and sexual activity — in communities, firms, hospitals, group medical practices, and schools (for example, Murray and Blitstein, 2003; Ukoumunne et al.; 1999; Siddiqui, Hedeker, Flay, and Hu, 1996; Murray and Short, 1995). The intra-class correlations for these clusters and outcomes are much smaller than those for measures of student achievement in schools and range from about 0.01 to 0.05.

The overall pattern of findings across categories of outcomes in Table 1 is thus consistent with these findings from prior research. However, values in the table for school-level intra-class correlations for academic outcomes are generally smaller than those observed by others. This may reflect two factors. First, findings in Table 1 are from three-level analyses, and those from most past research are from two-level analyses. As demonstrated in Part IV, estimates of school-level intra-class correlations from a three-level analysis are systematically smaller than those from a two-level analysis of the same date. Second, the samples of schools for CLIMBERs and the SBPP may be more homogenous than those for entire school districts that have been used for most related prior research (for example, Hedges and Hedberg, 2007; Bloom, Richburg-Hayes, and Black, 2007).

Explanatory Power of Covariates

CLIMBERs collected baseline data on reading pretests to use as a covariate. These data were obtained for individual students, but because there was so much student mobility (and thus attrition) during the school year between the pretest and posttest, this information was aggregated to the school level for use as a covariate. This was accomplished by computing the mean value of individual student pretest scores for each school. CLIMBERs also collected school-level demographic information, such as average student age, gender, ethnicity, and eligibility for free or reduced-price lunch, to use

as covariates. The SBPP study collected baseline student-level pretest information, plus student-level demographic information.

The last three columns of Table 1 present the estimated R-squared value or proportion of variance "explained" by covariates for each outcome measure. In each case the best possible combination of covariates (those with the most explanatory power) was used. For CLIMBERs, only school-level covariates were used, whereas for the SBPP, student-level covariates were used. No classroom-level covariates were used.

Consider first the findings for academic outcomes. All classroom-level and student-level R-squared values equal zero for outcome measures from CLIMBERs, because only school-level covariates could be used for these outcomes. School-level covariates do not vary across classrooms within schools or across students within classrooms, so they cannot co-vary with classroom-level or student-level outcomes. Consequently, they have *zero* explanatory power for classroom or student variation. On the other hand, R-squared values from the SBPP (which used student-level pretests and demographic information as covariates) for SAT 9 math and reading scores are substantial at both the classroom level (0.627 and 0.880) and the student level (0.482 and 0.510). For academic outcome measures from both studies, R-squared values for school-level variation ranged from 0.346 to 1. The one exception was Elision, for which an R-squared value could not be estimated because its unconditional school-level variance was zero.

For other outcomes in the table, we were able to calculate R-squared values only for student-level demographic covariates. Adding students' age, gender, ethnicity, and free or reduced-price lunch status reduced the student-level variance by very little, however. These covariates also reduced the classroom-level variance by very little. On the other hand, they reduced school-level variances appreciably for several outcome measures. This is an important finding for the design of group-randomized studies, because the school-level variance component is usually the primary factor that determines the required sample size.

A number of the R-squared values reported in Table 1 are negative. This could be due to estimation error, which can occur when the estimated unconditional variance is close to zero. In this case, a small amount of estimation error can produce an estimated conditional variance component that is larger than its unconditional counterpart, thus producing a negative value for R-squared. It is also possible that, after controlling for level 1 covariates, the level 2 variance actually increased, which would lead to a negative value of the R-squared.

Lastly, note that several R-squared values in the table are equal to one, which implies that the covariate or covariates involved explain all of a variance component. This occurs only for school-level variance components that are very close to zero without covariates. Hence, there is not much variance at this level for covariates to explain.

In addition to estimating R-squared values for the best possible model for each outcome measure, we also investigated the explanatory power of different combinations of covariates. Table 2 presents results of analyses for pretests alone, demographic characteristics alone, and pretests plus demographic characteristics together. Findings are reported for the subset of outcomes that have both a preprogram measure (pretest) and demographic information.

For the first four outcomes in the table, only school-level covariates are available, and thus only R-squared values for the school-level variance component are nonzero. For these outcomes, it is clear that the explanatory power of school-level demographic variables is much less than that of school-level pretest measures. Furthermore, the added value of combining these two types of covariates is limited.

The remainder of the table presents results for outcome measures that have student-level covariates. For reading and math test scores, demographic covariates provide slightly more explanatory power than pretests at the school and classroom levels, but the reverse is true at the student level. Adding demographic covariates to pretests does not consistently improve explanatory power at any of the three levels. There is a similar pattern of findings for program participation, attendance, and tardiness, although their R-squared values are smaller than those for academic outcomes.

Using Parameter Estimates to Compute Minimum Detectable Effect Sizes

The payoff from collecting data about intra-class correlations and R-squared values is the ability to use this information to estimate MDES for alternative sample designs. Table 3 illustrates the results of doing so based on the intra-class correlations and R-squared values reported in Table 1.

The first column of Table 3 reports the MDES of the original sample used for the present analysis (Appendix C reports the sample size and structure for each outcome measure). Recall that for the CLIMBERs data there were 430 students from 47 classrooms in 23 schools in one school district. This represents about 9 students per classroom and 2 classrooms per school. Given this sample and the estimated intra-class cor-

Table 2 Estimated R-Squared Values from Models with Different Sets of Covariates

Outcome	Pretest			Demographics[a]			Pretest + Demographics		
	School	Class	Student	School	Class	Student	School	Class	Student
Academic Outcomes									
Print Awareness (CLIMBERs)	0.580	0.000	0.000	0.200	0.000	0.000	0.889	0.000	0.000
Blending (CLIMBERs)	0.346	0.000	0.000	-0.053	0.000	0.000	-0.010	0.000	0.000
Elision (CLIMBERs)	n.e.	0.000	0.000	n.e.	0.000	0.000	n.e	0.000	0.000
Expressive Vocabulary (CLIMBERs)	1.000	0.000	0.000	0.394	0.000	0.000	1.000	0.000	0.000
Stanford 9 total math scaled score	0.454	0.421	0.474	0.585	0.418	0.069	0.494	0.627	0.482
Stanford 9 total reading scaled score	0.808	0.820	0.503	0.875	0.196	0.066	0.840	0.880	0.510
Academic-Related Outcomes									
Breakfast participation (adjusted for attendance)	0.358	n.e.	0.289	0.217	n.e.	0.120	0.385	n.e.	0.320
Attendance	n.e.	0.499	0.311	n.e.	0.149	0.024	n.e.	0.525	0.311
Days tardy as a percentage of no.of school days enrolled	0.214	n.e.	0.195	0.113	n.e.	0.017	0.253	n.e.	0.217

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

[a]Demographic information includes age, ethnicity, gender, and eligibility for free/reduced lunch.

n.e.=not estimable.

n.a.=not available.

Table 3 Calculated Minimum Detectable Effect Size (MDES) from Three-Level Models

	Original Sample Structure (varies by outcome)	Hypothetical Sample Structure							
Number of Students Per Class		**5**	**5**	**5**	**5**	**25**	**25**	**25**	**25**
Number of Classes Per School		**2**	**2**	**4**	**4**	**2**	**2**	**4**	**4**
Number of Schools		**20**	**100**	**20**	**100**	**20**	**100**	**20**	**100**
Outcome									
Academic Outcomes									
Print Awareness[b] (CLIMBERs)	0.516	0.567	0.254	0.512	0.229	0.486	0.218	0.469	0.210
Blending[b](CLIMBERs)	0.476	0.541	0.242	0.472	0.211	0.433	0.194	0.412	0.184
Elision[b] (CLIMBERs)	0.357	0.446	0.200	0.316	0.141	0.287	0.128	0.203	0.091
Expressive Vocabulary[b](CLIMBERs)	0.372	0.453	0.202	0.320	0.143	0.313	0.140	0.221	0.099
Stanford 9 total math scaled score[a,c]	0.184	0.380	0.170	0.323	0.144	0.294	0.131	0.274	0.123
Stanford 9 total reading scaled score[a,c]	0.148	0.298	0.133	0.227	0.102	0.190	0.085	0.159	0.071
Academic-Related Outcomes									
Breakfast participation (adjusted for attendance)[c]	0.243	0.532	0.238	0.491	0.219	0.464	0.208	0.455	0.203
Attendance[c]	0.170	0.385	0.172	0.272	0.122	0.259	0.116	0.183	0.082
Emotional and Behavioral Outcomes									
Conners' ADHD Index[c]	0.198	0.454	0.203	0.324	0.145	0.309	0.138	0.222	0.099
Cognitive Problems/Inattention[c]	0.172	0.396	0.177	0.280	0.125	0.215	0.096	0.152	0.068
Health Outcomes									
Body Mass Index percentile[c]	0.166	0.395	0.177	0.279	0.125	0.177	0.079	0.125	0.056
At risk of overweight[c]	0.170	0.402	0.180	0.290	0.130	0.194	0.087	0.148	0.066

Sources: Where indicated, calculations are based on data from the CLIMBERs database; all other calculations are based on data from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the intra-class correlations were obtained from a three-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for each school district in the study sample.

[a]Baseline measure of the outcome variable is included as prior achievement measure in the model.

[b]Baseline measure of other academic outcomes is included as prior achievement measure in the model.

[c]Student-level demographic information (age, ethnicity, gender, eligibility for free/reduced lunch) is included in the model.

relations and R-squared values reported in Table 1, MDES range from about 0.37 to 0.52 standard deviations for the four CLIMBERs outcome measures.

For the SBPP data set, the sample varies from outcome to outcome because of missing data. In general, this dataset represents about 1,100 students from 230 classes in 110 schools (or about 5 students per classroom and 2 classrooms per school). The estimated MDES for its outcome measures range mainly from about 0.15 to 0.20 standard deviation.

The remaining columns in the table vary the sample size and structure while holding constant the estimated values of intra-class correlations and R-squared values (The findings in the table assume that half the schools are randomized to treatment status, and half are randomized to control status). This illustrates how to assess the implications for precision of alternative sample sizes and structures.

Columns two and three in the table illustrate for each outcome measure how a fivefold increase in the number of randomized schools from 20 to 100 reduces the minimum detectable size, given five students per classroom and two classrooms per school. For Print Awareness, doing so reduces the MDES from 0.567 standard deviations to 0.254 standard deviations.

By comparing the findings in columns two and five, one can examine the effect of a fivefold increase in the number of students per school produced by a fivefold increase in the number of students per classroom. For Print Awareness, this implies a reduction in the MDES from 0.567 standard deviations to 0.486 standard deviations.

Comparing the two preceding sets of results illustrates the well-known fact that a proportional increase in the number of schools (or more generally randomized groups) improves precision by far more than the same proportional increase in the number of students per school (For example, see Bloom, Richburg-Hayes, and Black, 2007).

Lastly, note how changing the number of classrooms and students per school influences precision, given a fixed total number of schools. This can be seen by comparing findings in columns two and four of the table. For example, given 20 randomized schools and doubling the number of classrooms per school 2 to 4 (and thereby doubling the number of students per school from 10 to 20) reduces the MDES for Print Awareness from 0.567 standard deviations to 0.512 standard deviations.

By making comparisons such as those just described, one can begin to assess the relative precision of alternative sample designs for specific outcome measures. And in this way a proposed research design can be developed and defended.

Part IV
Assessing Two-Level Designs for Three-Level Situations

The Methodological Problem

This section addresses the question: 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. As noted earlier, most of the empirical basis for planning such studies comprises information for the two-level variance structure of students clustered in schools. 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 often do not identify which students are in which classrooms — and adding such identifiers is difficult or costly, if not impossible to do — these studies are analyzed using two-level models that do not account explicitly for the clustering of students within classrooms. We demonstrate below that, even though this middle level of clustering is not accounted for *explicitly* in the design or analysis of many studies, it is actually accounted for *implicitly*.

The Statistical Issue

To help understand what is at stake, consider two alternative research designs for estimating the impacts of an educational intervention on student outcomes from a study that randomizes elementary schools in a large urban school district. Impacts are estimated by the observed differences in mean outcomes for the randomized treatment group and control group. Without loss of generality, we assume that the outcome variable has a standard deviation of 1.0. In this case, impact estimates represent standardized effect sizes. Also assume that the true variance structure comprises three levels, with students clustered in classrooms that are clustered in schools.

Design A uses a three-level variance model, which explicitly recognizes all three levels of the variance structure. The true school-level variance equals τ_A^2, which is the variance of mean outcomes across schools. The true classroom-level variance equals γ_A^2, which is the variance of classroom means within schools. The true 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.

Design B uses a two-level variance model, which recognizes only a variance for mean values of the outcome measure across schools, τ_B^2 and a variance 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. By ignoring the clustering of students within classrooms, 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 (MDES) for mean student outcomes, given designs A and B, without covariates or blocking:[12]

Design A

$$MDES_A = \frac{M_{J-2}}{\sqrt{P(1-P)}}\sqrt{\frac{\tau_A^2}{J}+\frac{\gamma_A^2}{JK}+\frac{\sigma_A^2}{JKN_A}} * \frac{1}{\sqrt{\tau_A^2+\gamma_A^2+\sigma_A^2}} \qquad (9)$$

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;
N_A= the harmonic mean number of students per classroom; and
K = the harmonic mean number of classrooms per school.

[12]As done throughout this paper, MDES are defined for a two-tail hypothesis test at the 0.05 level of statistical significance with 80 percent statistical power.

Design B

$$MDES_B = \frac{M_{J-2}}{\sqrt{P(1-P)}}\sqrt{\frac{\tau_B^2}{J} + \frac{\sigma_B^2}{JN_B}} * \frac{1}{\sqrt{\tau_B{}^2 + \sigma_B{}^2}} \quad (10)$$

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. 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). Their difference lies within the standard error of the impact estimator, which is represented by the square root of the sum of contributions of the variances from the different levels of the statistical model.

The central question to address when comparing these two expressions is: How do their *estimated* values compare if all three variances are estimated and used for Equation 9, but only the top- and bottom-level variances are estimated and used for Equation 10?

To help develop some intuition about 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 estimated variances under model B or:

$$\tau_A^2 + \gamma_A^2 + \sigma_A^2 = \tau_B^2 + \sigma_B^2 \quad (11)$$

Variance estimates for model B must thus "shift" some of the middle-level variance to the bottom level, the top level, or both levels. Appendix D proves that the following expressions represent this shifting:

$$E(\hat{\tau}_B^2) = \tau_A^2 + \left(\frac{N_A - 1}{N_A(K-1)}\right)\gamma_A^2 \quad (12)$$

$$E(\hat{\sigma}_B^2) = \sigma_A^2 + \left(\frac{N_A(K-1)}{N_A(K-1)}\right)\gamma_A^2 \quad (13)$$

Where:

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

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

Equations 12 and 13 indicate that part of the true classroom-level variance is shifted to the estimated school-level variance, while the remainder is shifted to the estimated student-level variance.

Intuitively, it is easy to see how part of the classroom-level variance "shifts down" to the estimated student-level variance. This occurs because part of the observed variance in outcomes across students within a school reflects classroom differences. Thus, when measuring the variation across students within schools and ignoring cross-classroom differences, part of these differences are included in the measure of student-level variance, $\hat{\sigma}^2_B$. Consequently, the estimated student-level variance for the two-level model $\hat{\sigma}^2_B$ exceeds that for the 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 means*. Thus, when decomposing the total observed variance in school means into the portion due to true variation across schools (the school-level variance) and the portion due to estimation error produced by within-school student variation, the two-level model overestimates the school-level variance. Consequently, the estimated school-level variance for the two-level model exceeds that for the three-level model.[13]

[13]Equation 12 indicates that less of the classroom-level variance is shifted to the estimated school-level variance as students per school (NAK) are clustered into fewer classrooms (K). This reflects how the clustering of students within classrooms inflates the true variability of within-school outcomes. Ignoring this clustering thus causes one to understate the within-school variability of outcomes by more when there are fewer classroom clusters, which, in turn, causes one to overstate the between-school variance accordingly.

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. As proved in Appendix D, using a two-level model to estimate the minimum detectable effect of a group-randomized research design will produce the same results as a three-level model, as long as the study used for planning purposes has the same sample allocation (the same number of students per classroom and the same number of classrooms per school) within schools as the study being designed.

The proofs of the preceding findings are based on a balanced sample for schools with the same number of classrooms and students per classroom. In addition, these proofs are for the expected values of the estimators being considered, not for specific estimates from a given sample. Furthermore these proofs do not consider situations in which covariates are used in the impact estimation model. The following section thus explores empirically how the findings apply to specific estimates from unbalanced samples, with varying numbers of classrooms per school and students per classroom, both with and without the use of covariates.

Empirical Findings

The first step in the empirical analysis was to estimate variances for each of the three levels in model A and for each of the two levels in model B from Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs) data and School Breakfast Pilot Project (SBPP) data.

Table 4 presents these variance estimates for 12 selected outcome measures. Corresponding findings for all other outcome measures are presented in Appendix C. Note that the variance estimates in Table 4 are not standardized and thus are reported in the original units of each outcome measure (squared). Because these variances are estimated without covariates, they are unconditional.

The first four columns in the table report nonstandardized unconditional variance estimates from the three-level model, plus the total variance across all students in each sample. The last three columns report nonstandardized unconditional variance estimates from the two-level model for students within schools, plus their total variance.

For example, the estimated three-level variances for Print Awareness (in the first row of the table) equal 32.2 at the school level, 1.7 at the classroom level, and 70.6 at the student level. Their sum equals 104.4, which is the total nonstandardized unconditional variance across all students from all classrooms in all schools in the CLIMBERs

Table 4 Three-Level vs. Two-Level Model Comparisons: *Nonstandardized* Unconditional Variance Components

	Nonstandardized Unconditional Variance Components						
	Three-Level Model				Two-Level Model		
Outcome	**School**	**Class**	**Student**	**Total**	**School**	**Student**	**Total**
Academic Outcomes							
Print Awareness (CLIMBERs)	32.2	1.7	70.6	104.4	33.2	71.4	104.6
Blending (CLIMBERs)	3.0	0.2	17.0	20.2	3.1	17.1	20.2
Elision (CLIMBERs)	0.0	1.1	14.8	15.9	0.5	15.4	15.9
Expressive Vocabulary (CLIMBERs)	19.8	32.5	306.2	358.5	38.2	321.1	359.3
Stanford 9 total math scaled score	115.1	36.4	1,273.2	1,424.7	131.4	1,293.2	1,424.6
Stanford 9 total reading scaled score	108.8	159.0	1,581.9	1,849.6	181.8	1,666.5	1,848.3
Academic-Related Outcomes							
Breakfast participation (adjusted for attendance)	193.4	0.0	745.9	939.2	193.5	745.8	939.3
Attendance	0.0	0.8	13.0	13.9	0.3	13.5	13.9
Emotional and Behavioral Outcomes							
Conners' ADHD Index	0.9	9.1	107.3	117.3	4.8	112.5	117.3
Cognitive Problems/Inattention	0.6	4.4	128.8	133.9	2.8	131.2	133.9
Health Outcomes							
Body Mass Index percentile	0.0	0.0	784.4	784.4	0.0	784.4	784.4
At risk of overweight	0.0	0.0	0.2	0.2	0.0	0.2	0.2

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

sample for that outcome measure.[14] The corresponding two-level variances are 33.2 at the school level and 71.4 at the student level, which total 104.6.

The first thing to note about these findings is that the three-level variance estimates and the two-level variance estimates sum to almost exactly the same total. Their only difference is due to estimation error. This finding holds for every measure in Table 4 and for every other measure in Appendix C. It reflects the fact that the three-level estimation model and the two-level estimation model start with the same total variance across all students.

The second thing to note about the 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. Hence, both of these estimated variances are larger for the two-level model than for the three-level model. This difference is not pronounced for the first outcome in the table, Print Awareness, because its classroom-level variance is small relative to those for the other two levels. The difference is more pronounced, however, for expressive vocabulary, because its classroom-level variance is appreciably larger relative to its other two variances. Its estimated school-level variance is 19.8 for the three-level model, versus 38.2 for the two-level model. And its estimated student-level variance is 306.2 for the three-level model, versus 321.1 for the two-level model.

Table 5 reports standardized estimates of the variances reported in Table 4, such that the three-level variance estimates for each outcome measure sum to a value of one, and the two-level variance estimates sum to a value of one. Consequently, the pattern of findings described above for Table 4 is also visible in Table 5. In addition, the results in Table 5 for school variances and classroom variances in the three-level model and for schools in the two-level model represent intra-class correlations.

Consider the findings for Print Awareness. In Table 5, the standardized variance for schools in the three-level analysis equals 0.308. In other words, the school-level intra-class correlation equals 0.308. This means that about 31 percent of the total variation across all students from all schools in the analysis sample is estimated to be due to differences in mean outcomes across schools. The standardized variance for classrooms in the three-level analysis equals 0.016. In other words, the classroom-level intra-class correlation equals 0.016. This means that approximately 2 percent of the total variation across all students from all schools in the analysis sample is estimated to be due to differences in mean outcomes for classrooms within schools. The remaining part of the

[14]As was shown in Part II, these variances were estimated using a statistical model that removes all existing differences between treatment and control groups. In addition, for the SBPP, the model removes all differences among the six participating school districts. Hence, all estimates represent within-district variances in the absence of treatment.

Table 5 Three-Level vs. Two-Level Model Comparisons: *Standardized* Unconditional Variance Components

	***Standardized* Unconditional Variance Components**				
	Three-Level Model			**Two-Level Model**	
Outcome	**School**	**Class**	**Student**	**School**	**Student**
Academic Outcomes					
Print Awareness (CLIMBERs)	0.308	0.016	0.676	0.318	0.682
Blending (CLIMBERs)	0.149	0.011	0.840	0.155	0.845
Elision (CLIMBERs)	0.000	0.068	0.932	0.032	0.968
Expressive Vocabulary (CLIMBERs)	0.055	0.091	0.854	0.106	0.894
Stanford 9 total math scaled score	0.081	0.026	0.894	0.092	0.908
Stanford 9 total reading scaled score	0.059	0.086	0.855	0.098	0.902
Academic-Related Outcomes					
Breakfast participation (adjusted for attendance)	0.206	0.000	0.794	0.206	0.794
Attendance	0.000	0.060	0.940	0.023	0.977
Emotional and Behavioral Outcomes					
Conners' ADHD Index	0.008	0.078	0.915	0.041	0.959
Cognitive Problems/Inattention	0.005	0.033	0.962	0.021	0.979
Health Outcomes					
Body Mass Index percentile	0.000	0.000	1.000	0.000	1.000
At risk of overweight	0.006	0.000	0.994	0.006	0.994

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

total variation is due to differences in outcomes for students within classrooms within schools.

The preceding findings demonstrate that none of the total variation in a three-level variance structure is "lost" when the middle level is not accounted for explicitly. The findings also demonstrate that not accounting for the middle level explicitly causes the estimated variances for both the top level and bottom level to increase.

Table 6 uses the standardized variance estimates from Table 5, plus the number of students, classrooms, and schools in the analysis sample for each outcome measure, to compute the MDES for that measure, given its original sample. Equation 9 was used to compute MDES for three-level analyses, and Equation 10 was used for two-level analyses.

To see how this was done, consider yet again the findings for Print Awareness. Given 23 schools with 47 classrooms and 430 students in the sample, plus the three-level standardized unconditional variance estimates of 0.308, 0.016 and 0.676 for schools, classrooms, and students, respectively (from Table 5), an unconditional MDES of 0.735 was computed using Equation 9. Similarly, given 23 schools with 430 students and the two-level standardized unconditional variance estimates of 0.318 and 0.682 for schools and students, respectively (from Table 5), an unconditional minimum detectable size of 0.737 was computed using Equation 10.[15]

These findings indicate that, whether the study had been planned using a two-level analysis or a three-level analysis, as long as the within-school allocation of classrooms and students is the same, the same statistical precision would have been predicted for Print Awareness, given the original sample. The same conclusion holds for all other outcome measures that were examined.

When assessing these findings, it is useful to examine the range of different relationships that exist among variances at different levels. For example, Print Awareness has a small proportion of its total variance at the classroom level (0.016) and a large proportion at the school level (0.308). Expressive Vocabulary has more of its total variance at the classroom level (0.091) and less at the school level 0.055). The Conners' ADHD Index has 0.078 of its total variance at the classroom level and 0.008 at the school level. Hence, the different outcome measures in the present analysis represent considerable diversity of variance structure. This suggests that the consistent relationship observed between three-level analyses and two-level analyses is not limited to a single idiosyncratic variance structure.

[15]These findings assume that half the schools are randomized to treatment, and half are randomized to control status.

Table 6 Three-Level vs. Two-Level Model Comparisons: Minimum Detectable Effect Size

	Minimum Detectable Effect Size			
	Three-Level Model		Two-Level Model	
Outcome	Unconditional	Conditional	Unconditional	Conditional
Academic Outcomes				
Print Awareness[b] (CLIMBERs)	0.735	0.521	0.737	0.526
Blending[b](CLIMBERs)	0.559	0.484	0.560	0.486
Elision[b] (CLIMBERs)	0.373	0.373	0.374	0.302
Expressive Vocabulary[b](CLIMBERs)	0.482	0.386	0.495	0.311
Stanford 9 total math scaled score[a,c]	0.259	0.184	0.259	0.184
Stanford 9 total reading scaled score[a,c]	0.261	0.148	0.264	0.150
Academic-Related Outcomes				
Breakfast participation (adjusted for attendance)[c]	0.305	0.243	0.305	0.243
Attendance[c]	0.211	0.170	0.210	0.163
Emotional and Behavioral Outcomes				
Conners' ADHD Index[c]	0.203	0.198	0.202	0.202
Cognitive Problems/Inattention[c]	0.186	0.172	0.187	0.164
Health Outcomes				
Body Mass Index percentile[c]	0.167	0.166	0.167	0.166
At risk of overweight[c]	0.172	0.170	0.172	0.170

Sources: Where indicated, calculations are based on data from the CLIMBERs database; all other calculations are based on data from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the intra-class correlations were obtained from a three-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for each school district in the study sample.

[a]Baseline measure of the outcome variable is included as prior achievement measure in the model.

[b]Baseline measure of other academic outcomes is included as prior achievement measure in the model.

[c]Student-level demographic information (age, ethnicity, gender, eligibility for free/reduced lunch) is included in the model.

The findings in Table 6 labeled *conditional* MDES take comparisons of two- and three-level analyses a step further by accounting for covariates. This is accomplished by including covariates in the models used to estimate multilevel variances and subsequently estimating the value of R-squared for each variance. Based on these estimated R-squared values and the original unconditional variances, it is possible to estimate the MDES for the original sample given available covariates.

The findings illustrate that MDES computed from a two-level analysis with covariates are almost identical to those computed from a three-level analysis with the same data and covariates. This can be seen by comparing the conditional MDES from a three-level analysis for a given outcome measure with its counterpart from a two-level analysis. To the extent that these results are similar to each other, using a two-level analysis, which does not account explicitly for the middle level of a three-level situation, does not produce misleading results when covariates are used. This is the case for all outcome measures that were examined. For example, the conditional MDES for Print Awareness is estimated to be 0.521 from a three-level analysis and 0.526 from a two-level analysis.

Table 6 compares only the estimated precision of two-level and three-level analyses for the original sample from which multilevel variances are estimated. These findings do not necessarily extrapolate to the typical situation in practice, where multilevel variances and R-squared values are computed from data for an existing study and then used to design a future study with a different sample size and 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.

Table 7 reports such findings. Columns one and two report three-level and two-level unconditional MDES for the original sample for each outcome measure (which are also reported in Table 6). Columns three and four report corresponding findings after doubling the number of classrooms per school but holding constant the number of students per classroom. Columns five and six report corresponding findings after doubling the number of students per classrooms but holding constant the number of classrooms per school. This provides three different comparisons of three-level versus two-level estimates of statistical precision. And each comparison represents a markedly different ratio of students to classrooms to schools. In all cases, the estimated MDES for three-level and two-level analyses are essentially the same.

Up until now our discussion has been focused on the variance components and MDES of two- versus three-level models. An additional question is whether or not the point estimate and standard errors on a treatment indicator included at the school level remain the same, whether or not a two- or three-level model is estimated. This question is particularly important, since in many instances researchers are not able to explicitly

Table 7 Minimum Detectable Effect Sizes for Alternative Sample Structures

Outcome	Minimum Detectable Effect Size Without Covariates					
	Original Sample Structure		Double Number of Classes		Double Number of Students	
	3-Level Model	2-Level Model	3-Level Model	2-Level Model	3-Level Model	2-Level Model
Academic Outcomes						
Print Awareness (CLIMBERs)	0.735	0.737	0.728	0.732	0.711	0.713
Blending (CLIMBERs)	0.559	0.560	0.545	0.555	0.516	0.523
Elision (CLIMBERs)	0.373	0.374	0.313	0.351	0.294	0.290
Expressive Vocabulary (CLIMBERs)	0.482	0.495	0.431	0.487	0.429	0.448
Stanford 9 total math scaled score	0.259	0.259	0.255	0.259	0.221	0.221
Stanford 9 total reading scaled score	0.261	0.264	0.248	0.264	0.225	0.226
Academic-Related Outcomes						
Breakfast participation (adjusted for attendance)	0.305	0.305	0.305	0.305	0.282	0.282
Attendance	0.211	0.210	0.201	0.210	0.162	0.159
Emotional and Behavioral Outcomes						
Conners' ADHD Index	0.203	0.202	0.188	0.202	0.165	0.162
Cognitive Problems/Inattention	0.186	0.187	0.180	0.187	0.143	0.143
Health Outcomes						
Body Mass Index percentile	0.167	0.167	0.167	0.167	0.118	0.118
At risk of overweight	0.172	0.172	0.172	0.172	0.125	0.125

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

link students to classes within schools and have no choice but to estimate a two-level model, despite the three-level structure of the data.

It can be shown that estimating a three-level model using Ordinary Least Squares (OLS) (that is, ignoring the nested nature of the data entirely) will provide unbiased estimates of impacts but will not be efficient, because the standard errors do not account for the nested nature of the data (Cheong, Fotiu, and Raudenbush, 2001). On the other hand, as shown in Appendix D, using feasible generalized least squares, which accounts for the nested nature of the data, will provide consistent and asymptotically efficient estimates for a three-level model if the sample size is large enough.[16] The question here is whether we can obtain consistent estimates of program impact if we misspecify the model by ignoring the second level of nesting, and whether the estimates will be asymptotically efficient.

The proof in Appendix D shows that for balanced samples (that is, the same number of students per classroom and the same number of classrooms per school) with no covariates at the student or classroom level, you will obtain identical estimates of program impact and identical standard errors, whether you explicitly account for the second level of nesting or not. These proofs are based on data that are balanced, which is rarely the case in practice, and do not consider situations in which covariates are included at the student or classroom level. In addition, these proofs are for the expected values of the estimators being considered, not for specific estimates from a given sample. *Therefore, Tables 8 and 9 show empirically how the estimates vary if we introduce unbalanced designs and/or covariates at lower levels of the model.*

Table 8 shows point estimates and standard errors on a school-level treatment indicator estimated using both a two- and three-level model, for selected outcomes from both the SBPP data and the CLIMBERs data. No covariates other than treatment status indicator and indicators for different sites were included in these models. While the point estimates are not identical, they are extremely close. This is true despite the range of outcomes explored and variations in the class-level Intra-Class Correlations (ICCs) for the various outcomes. The same is true of the standard errors of the estimates. Table 9 shows these same models, but with student-level covariates included.[17] The pattern of findings is the same, confirming empirically that the practice of using two-level models (students nested in schools), rather than three-level models (students in classrooms in schools), to estimate school-level treatment effects does not lead to misleading findings.

[16]The crucial sample size is the number of level-3 units, and it depends upon how large the variation is among these units and how unbalanced the data are. Cheong, Fotiu, and Raudenbush (2001) provide a simulations study to probe the adequacy of level-3 sample sizes.

[17]Students' age, gender, ethnicity, and eligibility for free/reduced-price lunch are included as covariates in all SBPP outcomes. For math and reading scores, breakfast participation, and attendance, students' preprogram measures of the outcome variable were also included in the regression.

Table 8 Three-Level vs. Two-Level Model Comparisons: Impact Estimates with No Covariates

	Impact Estimates			
	Three-Level Model		**Two-Level Model**	
Outcome	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Academic Outcomes				
Print Awareness (CLIMBERs)	-0.444	2.545	-0.442	2.554
Blending (CLIMBERs)	1.848	0.846	1.851	0.848
Elision (CLIMBERs)	-0.193	0.488	-0.223	0.488
Expressive Vocabulary (CLIMBERs)	3.045	3.067	2.992	3.142
Stanford 9 total math scaled score	4.254	3.514	4.263	3.514
Stanford 9 total reading scaled score	4.959	4.061	5.060	4.074
Academic-Related Outcomes				
Breakfast participation (adjusted for attendance)	15.480	3.426	15.479	3.426
Attendance	-0.354	0.269	-0.346	0.266
Emotional and Behavioral Outcomes				
Conners' ADHD Index	-0.442	0.786	-0.418	0.783
Cognitive Problems/Inattention	-0.499	0.770	-0.483	0.773
Health Outcomes				
Body Mass Index percentile	1.651	1.657	1.651	1.657
At risk of overweight	0.023	0.028	0.023	0.028

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

Even when the data are substantially unbalanced, the estimator based on the mis-specified model will still be consistent (See Equation D.29 in Appendix D). However, estimated standard errors will not generally be correct. To obtain accurate standard errors, one may use Huber-White corrected standard errors clustered at the school level, as long as the number of schools is not too small (Raudenbush and Bryk, 2002). Cheong, Fotiu, and Raudenbush (2001) have conducted a simulation study of the behavior of these robust standard errors when the three-level model is mis-specified as a two-level model, just as in our study.

Interpretation

This section demonstrates that the current practice of not accounting explicitly for the middle level of a three-level situation when planning a group-randomized study is a reasonable concession to reality.

Table 9 Three-Level vs. Two-Level Model Comparisons: Impact Estimates with Student-Level Covariates

	Impact Estimates			
	Three-Level Model		**Two-Level Model**	
Outcome	**Coefficient**	**S.E.**	**Coefficient**	**S.E.**
Academic Outcomes				
Print Awareness (CLIMBERs)	1.114	1.285	1.111	1.297
Blending (CLIMBERs)	2.034	0.837	2.029	0.838
Elision (CLIMBERs)	0.022	0.464	0.019	0.426
Expressive Vocabulary (CLIMBERs)	4.487	2.361	4.472	2.340
Stanford 9 total math scaled score	-2.002	2.506	-2.015	2.502
Stanford 9 total reading scaled score	-1.507	2.299	-1.488	2.316
Academic-Related Outcomes				
Breakfast participation (adjusted for attendance)	16.778	2.733	16.778	2.734
Attendance	-0.388	0.216	-0.385	0.204
Emotional and Behavioral Outcomes				
Conners' ADHD Index	-0.323	0.768	-0.287	0.769
Cognitive Problems/Inattention	-0.286	0.711	-0.281	0.676
Health Outcomes				
Body Mass Index percentile	1.813	1.655	1.813	1.655
At risk of overweight	0.026	0.028	0.026	0.028

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

Students' age, gender, ethnicity, and eligibility for free/reduced price lunch are included as covariates in all SBPP outcomes. For math and reading scores, breakfast participation, and attendance, students' preprogram measures of the outcome variable were also included in the regression.

In closing, three further points should be noted. The first point is one of clarification. What this section has been discussing is not accounting explicitly for the classroom variance component by subdividing total student variance into two components (for schools and students within schools), instead of three components (for schools, classrooms within schools, and students within classrooms). By doing so, part of the classroom-level variance component shows-up in *estimates* of the school-level variance component, and part shows-up in *estimates* of the student-level variance component. In this way, the classroom-level variance remains in the analysis of minimum detectable effects. Consequently, it is accounted for indirectly, even though it is not included directly.

What the present section does *not* discuss is using a three-level analysis to compute variance components for schools, classrooms, and students, and then not including the classroom component in computations of minimum detectable effects. Doing so will understate the true precision of a sample design whenever the classroom variance is nonzero.

Secondly, our analyses do not include classroom-level covariates, such as teacher characteristics. If such covariates were to be collected and used in an impact analysis, then one would ideally account for them while planning a study. This can be done, however, only if appropriate information on unconditional intra-class correlations and R-squared values were available for all three levels, which to date has occurred only in rare instances. Whether or not this is an important problem is not clear, however. On one hand, the best covariates for student outcomes (the primary focus of most educational evaluations) are past values of the outcome being used for the impact analysis ("pretests"). If such information is at the student level, it can have substantial explanatory power for all three levels of an analysis. Thus, the added predictive power of classroom-level covariates might be modest.

Finally, the preceding findings are contingent on the fact that the sample allocations (number of students per class and number of classrooms per school) are held constant between the study used for planning purposes and the study being designed.

Part V
Accounting for Uncertainty About Intra-Class Correlations

As noted throughout this paper, researchers rely heavily on estimates of intra-class correlations to design group-randomized studies because these parameters have a major effect on the required sample size. For example, in a two-level analysis, assuming an intra-class correlation of 0.15 instead of 0.05 can, under certain circumstances, almost double the number of clusters needed to obtain a given level of precision.

Several factors should be considered when determining how much confidence to place in an estimate of an intra-class correlation for use in planning such studies. First, one must consider how similar the planned study sample will be to the sample used to estimate the intra-class correlation. For example, estimates of intra-class correlations from a small rural community may not be appropriate for planning a study that will take place in a large urban school district. Similarly, estimates of intra-class correlations based on a common outcome measure will most likely provide a better planning guide than those for different outcome measures.

Another important and often overlooked consideration when assessing the appropriateness of an estimated intra-class correlation for planning a study is the statistical uncertainty that exists about the estimate. This uncertainty depends on the number of clusters and subjects per cluster in the estimation sample. In addition, it depends on the true value of the intra-class correlation.

Taking this uncertainty into account is especially important when a researcher might otherwise have confidence in an estimated intra-class correlation because it comes from the same population and is based on the same outcome measure as that for the study being planned. For example, a researcher using an estimated intra-class correlation from a small-scale pilot study to plan a large-scale impact evaluation should consider carefully the uncertainty that exists about the estimate of the intra-class correlation.

This section of the paper considers how to measure and interpret the uncertainty that exists about intra-class correlations for two-level research designs.[18] Two-level designs are considered because most studies to date have employed them and because the

[18] A similar problem of uncertainty arises when using estimated values of R-squared to plan a group-randomized study. This issue is beyond the scope of this paper, however.

statistical properties of their intra-class correlations are relatively well understood. The discussion of uncertainty proceeds as follows: (1) it describes how standard errors and confidence intervals can be calculated for estimates of two-level intra-class correlations; (2) it examines the factors that influence these indicators of uncertainty; and (3) it illustrates their implications for findings from the Chicago Literary Initiative: Making Better Early Readers study (CLIMBERs) and the School Breakfast Pilot Project (SBPP) study.

Estimating Uncertainty for Intra-Class Correlations

According to Siddiqui, Hedeker, Flay, and Hu (1996), the variance of an estimated intra-class correlation for a two-level model was originally derived by Fisher (1925) and can be estimated as follows[19]:

$$Var(\hat{\rho}) = \frac{2(1-\hat{\rho})^2[1+(N-1)\hat{\rho}]^2}{N(N-1)J} \qquad (14)$$

where:

$\hat{\rho}$ = the estimated intra-class correlation;
N = the harmonic mean number of individuals per cluster; and
J = the total number of clusters.

The standard error of the estimated intra-class correlation equals the square root of the expression in Equation 14. Note that this standard error assumes that all studies have the same true intra-class correlation and that the only variation that arises among their estimates is sampling error. In reality, the largest source of variation among studies may be differences in their true intra-class correlations. The estimates presented here cannot take this variation into account.

Table 10 illustrates how the standard error derived from Equation 14 varies with $\hat{\rho}$, N, and J. First, note that as the number of clusters (J) increases, the standard error of the estimated intra-class correlation decreases. In fact, Equation 14 implies that

[19]Equation 14 is subject to some debate. For example, Visscher (1998) argues that it is probably wrong because it takes an expression derived when ρ is known and substitutes an estimated value for ρ. In addition, variants of the formula replace *N* with *N-1* or *N-2*. However, as long as the clusters contain at least 10 individuals, these differences in formulation are not important. The above formulation is quite accurate as ρ becomes small and *N**J becomes large.

Table 10 Standard Error of the Estimated Intra-Class Correlation (ICC) Given the Estimated ICC, Cluster Size (N), and Number of Clusters (J)

	Students Per School =10		Students Per School=50	
Intra-Class Correlation	10 Schools	50 Schools	10 Schools	50 Schools
0.0	0.047	0.021	0.009	0.004
0.1	0.081	0.036	0.048	0.021
0.2	0.106	0.047	0.078	0.035
0.3	0.122	0.055	0.099	0.044
0.4	0.130	0.058	0.112	0.050
0.5	0.130	0.058	0.115	0.052
0.6	0.121	0.054	0.110	0.049
0.7	0.103	0.046	0.096	0.043
0.8	0.077	0.035	0.073	0.032
0.9	0.043	0.019	0.041	0.018

Source: Authors' calculation based on Equation 14.

the estimated standard error is inversely proportional to the square root of J. For example, with an estimated intra-class correlation of 0.5 and 10 individuals per cluster, theestimated standard error of the intra-class correlation decreases from 0.130 to 0.058 (by the square root of five) as the number of clusters quintuples from 10 to 50.

Second, note that as the number of individuals per cluster (N) increases, the standard error of the estimated intra-class correlation also decreases, although this relationship is more complex than that for the number of clusters. For example, with an intra-class correlation of 0.5 and a total of 10 clusters, the estimated standard error of the intra-class correlation decreases from 0.130 to 0.115 as the number of individuals per cluster quintuples from 10 to 50.

The preceding results illustrate that a proportional increase in the number of clusters reduces the standard error of the intra-class correlation by far more than the same proportional increase in the number of individuals per cluster. Hence, the relative influence of clusters and individuals on the uncertainty about estimates of intra-class correlations is similar to their relative influence on the precision of intervention effects from group-randomized studies.

Lastly, note that the standard error of an intra-class correlation decreases to a minimum as the value of the intra-class correlation approaches zero or one and increases to a maximum as the value of the intra-class correlation approaches 0.5. For example, with 10 clusters and 10 individuals per cluster, the estimated standard error of the intra-class correlation decreases from 0.130 to 0.081 or 0.043 as the value of the intra-class correlation changes from 0.5 to 0.1 or 0.9, respectively.

A confidence interval for an estimated intra-class correlation equals the point estimate (the actual estimated value), plus or minus a multiple of the estimated standard error. The multiple to use for this purpose is obtained from the t distribution for the confidence level specified and the number of degrees of freedom available for estimating the cluster-level variance component, τ^2.

For example, assume that an intra-class correlation was estimated from a sample of 50 clusters with 10 individuals per cluster, using a two-level model with no covariates. If the estimated intra-class correlation (the point estimate) were 0.20, then according to Table 8, the estimated standard error would be 0.047. With no covariates and no treatment indicator variable, the number of degrees of freedom for estimating τ^2 equals the number of clusters minus one (J-1). This implies 49 degrees of freedom for the present example. For a t distribution with 49 degrees of freedom, the 95 percent confidence interval would be $0.20 \pm 2.01*0.047$, which ranges from about 0.1 to 0.29. Consequently, there would be considerable uncertainty about the value of the intra-class correlation to use for planning the study.

One way to account for this uncertainty is to assess sample size requirements using the lower bound of the confidence interval, the point estimate of the intra-class correlation, and the upper bound of the confidence interval. The best single estimate of the sample size requirement is that based on the point estimate for the intra-class correlation. But given the existing uncertainty about this estimate, it would be prudent to plan for a sample that is somewhat larger than that implied by the point estimate. Doing so would help guard against the possibility of underestimating the intra-class correlation and thus "undersizing" the study sample, thereby underpowering the study estimators.

Uncertainty about Intra-Class Correlations for This Paper

Table 11 presents point estimates, estimated standard errors, and 95 percent confidence intervals for two-level unconditional intra-class correlations obtained from data for CLIMBERs and the SBPP study. (Equation 12 was used to estimate standard errors). The first column in the table lists the estimated intra-class correlation for each outcome measure; the second column lists the estimated standard error of the intra-class correlation; and the final two columns list the corresponding 95 percent confidence interval.

These findings illustrate that the relatively small size of the CLIMBERs sample (with 430 students from only 23 schools) leaves considerable uncertainty about estimates of intra-class correlations. For example, the confidence interval for Print Awareness, the measure with the largest estimated intra-class correlation, ranges from 0.222 to

Table 11 Standard Errors and 95 Percent Confidence Intervals for the Estimated Intra-Class Correlations (ICC) from Unconditional Two-Level Models

Outcomes	Intra-Class Correlation (ICC)	Standard Error of ICC	95% Confidence Interval of ICC	
Academic Outcomes				
Print Awareness (CLIMBERs)	0.318	0.050	0.222	0.418
Blending (CLIMBERs)	0.155	0.035	0.092	0.228
Elision (CLIMBERs)	0.032	0.015	0.001	0.059
Expressive Vocabulary (CLIMBERs)	0.106	0.028	0.055	0.165
Stanford 9 total math scaled score	0.092	0.009	0.074	0.110
Stanford 9 total reading scaled score	0.098	0.010	0.079	0.117
Academic-Related Outcomes				
Breakfast participation (adjusted for attendance)	0.206	0.017	0.173	0.239
Attendance	0.023	0.003	0.017	0.029
Emotional and Behavioral Outcomes				
Conners' ADHD Index	0.041	0.004	0.032	0.050
Cognitive Problems/Inattention	0.021	0.003	0.015	0.026
Health Outcomes				
Body Mass Index percentile	0.000	0.001	-0.002	0.002
At risk of overweight	0.006	0.001	0.004	0.009

Sources: Where indicated, data are from CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

0.418; that for Elision, the measure with the smallest estimated intra-class correlation, ranges from 0.001 to 0.059. This means that the "true" value of the intra-class correlation for Print Awareness is equally likely to be anywhere between 0.222 and 0.418, and the true value of the intra-class correlation for Elision is equally likely to be anywhere between 0.001 and 0.059.

A comparison of these findings for the two outcome measures also illustrates how the magnitude of the underlying intra-class correlation affects the width of the confidence interval, given a constant sample size and configuration. The width of the confidence interval for Print Awareness (with a point estimate of 0.316) is 0.196, whereas the width of the confidence interval for Elision (with a point estimate of 0.032) is only 0.058.

Intra-class correlations from the SBPP were based on data for 800 to 1,000 students from approximately 100 schools or 8 to 10 students per school. (Samples vary across outcome measures due to missing data.) Hence, the uncertainty about these estimates is less than for estimates from the CLIMBERs sample. For participation in the school breakfast program, the SBPP measure with the largest estimated intra-class correlation, the confidence interval is 0.173 to 0.239. For "at risk of overweight," the SBPP measure with the smallest nonzero estimated intra-class correlation, the confidence interval is 0.004 to 0.009. A comparison of results for these two outcome measures also illustrates how the magnitude of the intra-class correlation affects the width of its confidence interval, given a constant sample size.

Implications of Uncertainty for Sample Design

Table 12 illustrates the implications of the preceding uncertainty for designing a group-randomized study. The first column in the table lists the predicted minimum detectable effect size (MDES) for an illustrative research design, given the *lower bound* of the confidence interval of the intra-class correlation for each outcome measure in Table 9. The second column presents corresponding results for the *point estimate* of the intra-class correlation, and the third column presents corresponding results for the *upper bound* of its confidence interval. The research design assumes 50 schools, with half randomized to treatment, 40 students per school, and use of the best-predicting covariates for each outcome measure (those used for Tables 1 and 3).

Note that the width of confidence intervals for MDES varies substantially across outcome measures in accord with the estimated standard errors for intra-class correlations. The width of this interval represents the degree of uncertainty that exists about the likely precision of impact estimates for the assumed research design. For example, the

Table 12 Minimum Detectable Effect Sizes (MDES) Associated with 95 Percent Confidence Intervals of the Estimated Intra-Class Correlation (ICC), from Two-Level Model with Covariates

Outcomes	MDES associated with 95% Confidence Interval of ICC		
	Lower Bound	Point Estimate	Upper Bound
Academic Outcomes			
Print Awareness[b] (CLIMBERs)	0.186	0.207	0.226
Blending[b](CLIMBERs)	0.230	0.284	0.329
Elision[b] (CLIMBERs)	0.126	0.146	0.163
Expressive Vocabulary[b](CLIMBERs)	0.133	0.140	0.146
Stanford 9 total math scaled score[a,c]	0.173	0.188	0.202
Stanford 9 total reading scaled score[a,c]	0.120	0.127	0.134
Academic-Related Outcomes			
Breakfast participation (adjusted for attendance)[c]	0.275	0.297	0.317
Attendance[c]	0.113	0.119	0.124
Emotional and Behavioral Outcomes			
Conners' ADHD Index[c]	0.184	0.197	0.209
Cognitive Problems/Inattention[c]	0.119	0.119	0.119
Health Outcomes			
Body Mass Index percentile[c]	0.121	0.125	0.129
At risk of overweight[c]	0.131	0.135	0.139

Sources: Where indicated, calculations are based on data from the CLIMBERs database; all other calculations are based on data from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the intra-class correlations were obtained from a three-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for each school district in the study sample.

[a]Baseline measure of the outcome variable is included as prior achievement measure in the model.

[b]Baseline measure of other academic outcomes is included as prior achievement measure in the model.

[c]Student-level demographic information (age, ethnicity, gender, eligibility for free/reduced lunch) is included in the model.

confidence interval of MDES for Blending (from CLIMBERs) is quite wide, ranging from 0.230 to 0.329 standard deviations. In contrast, the confidence interval of MDES for school breakfast participation (from the SBPP study) is much narrower, ranging from 0.275 to 0.317 standard deviation.

Table 13 moves the discussion of uncertainty a step further by translating the findings in Table 10 into their implications for the number of randomized schools needed to achieve a MDES of 0.25 standard deviations. The first column in the table assumes the lower bound of the confidence interval for each intra-class correlation, the second column assumes the point estimate, and the third column assumes the upper bound of the confidence interval. These findings provide a readily interpretable way to view the implications for research design of uncertainty about intra-class correlations.

Table 13 Number of Schools Needed for Minimum Detectable Effect Size (MDES) of 0.25

Outcomes	Number of Schools Needed for MDES = 0.25		
	ICC = Lower Bound	ICC = Point Estimate	Bound
Academic Outcomes			
Print Awareness[b] (CLIMBERs)	28	34	41
Blending[b](CLIMBERs)	42	64	86
Elision[b] (CLIMBERs)	13	17	21
Expressive Vocabulary[b](CLIMBERs)	14	16	17
Stanford 9 total math scaled score[a,c]	24	28	33
Stanford 9 total reading scaled score[a,c]	12	13	14
Academic-Related Outcomes			
Breakfast participation (adjusted for attendance)[c]	61	70	80
Attendance[c]	10	11	12
Emotional and Behavioral Outcomes			
Conners' ADHD Index[c]	27	31	35
Cognitive Problems/Inattention[c]	11	11	11
Health Outcomes			
Body Mass Index percentile[c]	12	13	13
At risk of overweight[c]	14	14	15

Sources: Where indicated, calculations are based on data from the CLIMBERs database; all other calculations are based on data from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the intra-class correlations were obtained from a three-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a three-level model of the outcome measure with and without student-level and school-level covariates where available. All analyses include an indicator variable distinguishing treatment and control groups; all analyses for outcomes from the SBPP database also include indicator variables for each school district in the study sample.

[a]Baseline measure of the outcome variable is included as prior achievement measure in the model.

[b]Baseline measure of other academic outcomes is included as prior achievement measure in the model.

[c]Student-level demographic information (age, ethnicity, gender, eligibility for free/reduced lunch) is included in the model.

Consider findings for the Blending measure from CLIMBERs. For this measure the projected number of required schools ranges from 42 to 86, with a point estimate of 64. This means that existing uncertainty about the value of the underlying intra-class correlation is so great that it is difficult to know how many schools are required. In contrast, findings for the Cognitive Problems/Inattention measure from the SBPP study reflect virtually no uncertainty (at least with respect to estimation error for the intra-class correlation) and thereby provide much clearer guidance for designing an experimental sample. Findings in the table suggest that this outcome would require about 11 randomized schools to achieve a MDES of 0.25 standard deviations.

Two main factors create the preceding differences in uncertainty about required sample sizes. First, the CLIMBERs sample has more schools from which to estimate an intra-class correlation than the SBPP sample (23 versus 100). Second, the value of the

intra-class correlation for Blending is larger than that for Cognitive Problems/Inattention. Both of these differences produce relatively more uncertainty about the intra-class correlation for Blending than for Cognitive Problems/Inattention.

Further Thoughts

This part of the paper has considered how to quantify the uncertainty that exists about intra-class correlations due to statistical estimation error and how to reflect this uncertainty in the sample size requirements of group-randomized studies. However, translating this information into sample size decisions is further complicated by several additional factors.

First, researchers must also consider the uncertainty that exists about estimates of the predictive power (R-squared) of covariates that will be used for a proposed impact analysis. Thus, future research should focus on how to quantify this uncertainty and how to obtain empirical information about its magnitude. A next logical step in this progression of knowledge would be to study the joint variation of estimates of intra-class correlations and R-squared values.

When this information becomes available, it will be possible to simulate how the joint uncertainty about these two planning parameters influences uncertainty about sample size requirements. With this information, a more fully-informed analysis of uncertainty about sample size requirements can be conducted as part of the planning process for group-randomized studies.

Nevertheless, there will always remain a need for researchers to translate *information* about uncertainty into *decisions* about sample size; and, to some extent, this step is fundamentally judgmental. Therefore, it must take into account the researchers' and research funders' attitudes toward risk, plus the cost structure of a proposed project. For example, other things being equal, a sample design for a high-profile study with high stakes attached to detecting intervention effects (if they exist) should tend to minimize the risk of inadequate precision. To do so would require erring on the side of a sample that might be larger than what is projected to be necessary.

In principle, one could develop a guide for such decisions by expanding the concept of confidence intervals to compute a probability distribution of required sample sizes for a given study design and desired level of precision. For example, one might simulate the required sample size at the 10^{th}, 20^{th}, 50^{th}, 80^{th}, and 90^{th} percentiles, given

whatever information is available to quantify existing uncertainty.[20] If such information could be obtained, then researchers could consciously decide how to manage their risks by choosing a sample size within this distribution. For example, in the previous example, where there would be considerable aversion to the risk of inadequate precision, a researcher might choose the projected sample size at the 80^{th} or 90^{th} percentile of the projected distribution. Of course, this would be possible only if the resources to do so were available.

[20]The 95 percent confidence intervals and point estimates in Table 11 represent the 5^{th}, 50^{th}, and 95^{th} percentiles of probability distributions for required sample sizes, based on estimated uncertainty about intra-class correlations.

Part VI
Afterword

The goal of this paper is to provide practical guidance for researchers who are designing studies that randomize groups to measure the impacts of interventions on children. The paper has proceeded by: (1) providing new empirical information about variance parameters that influence the precision of impact estimates; (2) examining the implications of planning group-randomized studies for three-level hierarchical situations, using empirical information obtained from estimating two-level models that omit the middle level; and (3) assessing the magnitude and implications of uncertainty that exists when estimating intra-class correlations for planning group-randomized studies. It is hoped that each of these small steps will move forward the current state of science of group-randomized studies.

Appendix A
Description of Outcome Measures

This Appendix presents detailed descriptions of the outcome measures discussed in the paper.[21]

Academic Outcomes

Academic outcomes are measured by conventional standardized achievement test scores. This paper includes the following measures:

Four subtests of the Preschool Comprehensive Test of Phonological and Print Processing[22] are used in CLIMBERs:

- **Print Awareness:** The Print Awareness subtest measures beginning knowledge about written language, for example, knowing what print looks like and how it works. Items measure whether children recognize individual letters, know what sounds letters make, and are able to differentiate words from pictures and other symbols.
- **Elision:** The Elision subtest tests a child's ability to segment spoken words into smaller parts, by deleting parts and then recalling the portion of the word. For example, "Say cup without saying /K/."
- **Blending:** The Blending subtest measures the child's ability to put sounds together to form words. For example, "What word do these sounds make: 't-oi' "?
- **Expressive Vocabulary:** The Expressive Vocabulary subtest measures the number of different vocabulary words an individual uses when he/she speaks or writes.

The Stanford 9 math and reading achievement tests (for the SBPP) measure total test scores in scaled score points.

[21]Discussions in this appendix are mainly based on Abt Associates Inc. and Promar (2005).

[22]The Pre-CTOPPP measures phonological skills development, which has been shown to be an important precursor to reading. The Pre-CTOPPP has not yet been published, and to date there is very little information about its psychometric properties — but it has been used widely with middle-income and low-income samples.

Academic-Related Outcomes

These measures, all from the SBPP, are not conventional direct measures of student academic performance. Rather, they measure children's behavior and other cognitive skills assessed in nonconventional ways. The following measures are included in this paper:

Participation, Attendance, and Tardiness

All districts have computerized attendance records that were used for this analysis. Attendance is defined as the number of days present at school, divided by the total number of school days the child was enrolled. Tardiness is defined as the number of days the student was late as a percentage of the number of school days the child was enrolled. Data on tardiness were not consistently available for all schools and districts. The amount of missing information is important to consider when interpreting the results.

Stimulus Discrimination

The Stimulus Discrimination measure (Detterman, 1988) has been used to assess the effects of breakfast on children's cognitive performance in several studies (Pollitt, Lewis, Garza, and Shulman, 1982/83; Pollitt and Matthews, 1998). The Stimulus Discrimination task is appropriate for children as young as 6 years of age. It is administered on a laptop computer and takes approximately 10 minutes to complete. It is appropriate for non-English speakers, as the entire task consists of attention to visual stimuli.

The Stimulus Discrimination measure is a modified match-to-sample test. The child is presented with six empty windows in a row slightly below the center of the screen. Centered above this row of windows is a probe window. When the child presses the space bar, the six windows each display a different stimulus. The probe window displays a probe identical to one of the stimulus items in the row below. The child needs to find the match to the probe in the bottom row, lift his/her finger, and touch the number key corresponding to the proper match.

When the child lifts his/her finger, all windows become empty. To view the items again, he/she has to press the space bar. The child views the stimulus display as long as desired, but the bar has to be pressed, or the display will show only empty windows.

After four practice trials, the child continues with the task until he/she responds correctly to 72 trials or completes 280 trials. Thus, the pacing of the task is entirely determined by the child. If, however, the child is not close to finishing 72 correct trials

after 15 minutes, the task is aborted so as to conserve time for the other components of the student interview. The variables used for analysis are as follows:

- Total Number of Trials: number of trials completed and number of trials incorrect;
- Average Viewing Time: total time of viewing stimuli averaged across all trials; and
- Average Trial Time: time in seconds from first press of space bar to answer; the total viewing and response time averaged across all trials.

Digit Span

The Digit Span task is a subtest of the Wechsler Intelligence Scales for Children III (WISC-III) (Wechsler, 1991). The WISC-III is a widely used standardized intelligence test with nationally representative norms. Subtests from the WISC-III are commonly used in developmental and neuropsychological research to assess child cognitive performance. It has previously been used to assess the effects of breakfast on child cognitive performance in several studies (Jacoby et al., 1996; Simeon and Grantham-McGregor, 1989). The Digit Span task is appropriate for children as young as 6 years of age and takes approximately five to seven minutes to administer.

The Digit Span task assesses short-term auditory memory and attentional abilities. A computer administration of the Digit Span was created for the SBPP evaluation. Through headphones, the child heard a recorded series of digits played by the computer. The child then repeated the series back to the tester, forwards in the first part of the task and backwards in the second part of the task.

On Digit Span Forwards, there were eight items, each with two trials of number series equal in length. The items increased in length until the child gave incorrect responses on both trials of any item or until the child reached the last trial, which is nine numbers on Digit Span Forwards and eight numbers on Digit Span Backwards. A total raw score of between 0 and 30 was possible on the Digit Span Task (Forwards + Backwards). This total raw score was then converted to a scaled score based on the child's age in years and months, to be used in the analysis.

Verbal Fluency

Verbal Fluency tasks are widely used to evaluate neuropsychological functioning in the areas of long-term verbal memory and retrieval and have been used in a number of studies of the effects of breakfast consumption on cognitive functioning (Simeon

and McGregor, 1989.) The Verbal Fluency task was considered age-appropriate for the children in the SBPP sample and takes approximately three minutes to administer.

Two scored trials of Verbal Fluency were administered following a practice trial to ensure that the child understood the task. The child was asked to name as many items as possible in two semantic categories ("animals" and "things to eat") in a period of 60 seconds each. The examiner recorded all of the child's answers, and the score equaled the total number of correct responses for each trial. As noise was a concern, headphones were worn during the task. Scores for both the Animals and Things to Eat trials were used for analysis, as well as a total of the two scores.

Children's Emotional and Behavioral Outcomes

The SBPP also provides a wide range of psychosocial and behavioral measures for young children, which are rare in studies in the education research field. This provides valuable information in terms of power calculation for design of studies including such outcomes.

Social/Emotional Functioning

In this study, social and emotional functioning was assessed through the Pediatric Symptom Checklist (PSC) (Murphy et al., 1998), included in the Parent Survey.

The PSC was developed for pediatricians to use as a screening tool for psychosocial problems. The version of the PSC used in this study is a 17-item questionnaire covering a broad range of children's social and emotional functioning, with the parent as the intended respondent (Gardner et al., 1999). The items are rated as "never," "sometimes," or "often" and are scored 0, 1, or 2, respectively. Item scores are summed, and the total score is recoded as a dichotomous variable. A score of 15 or higher is considered positive for psychosocial impairment. A score below 15 is negative. Examples of items include: "Feels sad, unhappy;" "Acts as if driven by a motor;" "Teases others;" and "Does not understand other people's feelings." Researchers indicate that nationally the prevalence of scores of 15 or higher is about 12 percent for middle class or "general" settings (http://psc.partners.org/psc_basic.htm). The mean of the 17-item questionnaire is about 8.[23]

[23]Communication with Michael Murphy.

Analyses of the PSC for this paper included a comparison of treatment and control students on total scores and on percentage of students considered psychosocially impaired.

Behavior

Behavioral measures used in the study come from two sources. The first one, Conners' Teacher Rating Scales-Revised, CTRS-R(s), is a part of a larger set of measures, the Conners' Rating Scales, which have long been used to assess psychopathology and behavior issues, such as problems with conduct, anxiety, and social functioning, as well as Attention Deficit Hyperactivity Disorder (ADHD) in children and adolescents (Conners, 2000). The CTRS-R(s) consists of 28 questions in which the teacher rates the child on a scale from 0 (not true at all/never or seldom) to 3 (very much true/very often or very frequent) and can be completed in an estimated five to 10 minutes. In scoring the CTRS-R(s), the 28 items are tallied within four constructs and are then scaled according to age and gender. They are as follows:

- Conners' ADHD Index: Identifies children "at risk" for ADHD[24]
- Cognitive Problems/Inattention: High scorers may have more academic difficulties than most individuals their age, have problems organizing their work, have difficulty completing tasks or schoolwork, and appear to have trouble concentrating on tasks that require sustained mental effort.
- Hyperactivity: High scorers have difficulty sitting still, feel more restless and impulsive than most individuals their age, and have the need to always be on the go.
- Oppositional: Individuals scoring high on this scale are more likely to break rules and have problems with persons in authority and are more easily annoyed and angered than most individuals their age.

The second source, the Effortful Control Scale, is comprised of a subset of questions from the Children's Behavior Questionnaire (CBQ), a highly differentiated assessment designed to measure temperament in children (Rothbart, Ahadi, and Evans, 2000). Two subscales, Ability to Focus (constructed from seven items) and Ability to Follow Instructions (constructed from six items), are used in this analysis.

[24]The four subscales were tested for internal consistency using Cronbach's alpha. Each subscale had high coefficients of reliability, ranging from .90 to .96, signaling that the individual items in each construct fit together very well in measuring the four latent constructs.

Health Outcomes

The SBPP also contains measures for student's basic health status, which are very important for studies in child development:

- Body Mass Index percentile
- The percentage of students "at risk of overweight"
- The percentage of students considered "overweight"
- Student's weight status
- Student's height
- Student's weight

Appendix B
Definition of the Multiplier M

The minimum detectable effect of a program impact estimator is a multiple M of its standard error (Bloom, 1995). Figure B.1 illustrates why this is the case. The bell-shaped curve on the left represents the t distribution, given that the true impact equals 0; this is the null hypothesis. For a positive-impact estimate to be statistically significant at the α level for a one-tailed test (or at the $\alpha/2$ level for a two-tailed test), it must fall to the right of the critical t-value, t_α (or $t_{\alpha/2}$), of this distribution. The bell-shaped curve on the right represents the t distribution, given that the impact equals the minimum detectable effect; this is the alternative hypothesis. For the impact estimator to detect the minimum detectable effect with probability 1-β (that is, to have a statistical power level of 1-β), the effect must lie a distance of $t_{1-\beta}$ to the right of the critical t-value of the alternative hypothesis and a distance of $t_\alpha + t_{1-\beta}$ (or $t_{\alpha/2} + t_{1-\beta}$) from the null hypothesis. Because t-values are expressed as multiples of the standard error of the impact estimator, the minimum detectable effect is also a multiple of the impact estimator. Thus, for a one-tailed test,

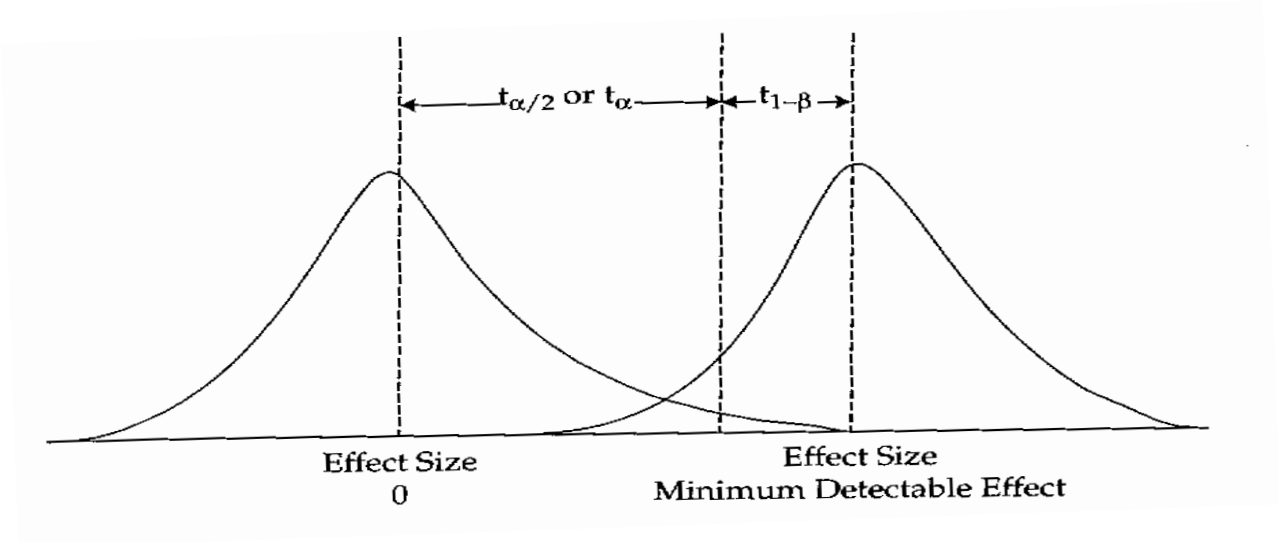
$$M = t_\alpha + t_{1-\beta} \qquad \text{(B.1)}$$

and for a two-tailed test:

$$M \approx t_{\alpha/2} + t_{1-\beta} \qquad \text{(B.2)}$$

The t-values in these expressions reflect the number of degrees of freedom available for the impact estimator, which for the full sample equals the number of clusters minus two (J-2). The multiplier for the full sample is thus referred to a M_{J-2}.

Figure B.1 The Minimum Detectable Effect Multiplier



One-Tailed Multiplier $M = t_{\alpha} + t_{1-\beta}$

Two-Tailed Multiplier $M \approx t_{\alpha/2} + t_{1-\beta}$

Source: Illustration by the authors.

Appendix C
Complete Set of Results for Three-Level vs. Two-Level Model Comparisons: Nonstandardized Unconditional Variance Components

Appendix Table C1 contains the following information for all outcome measures discussed in Appendix A: (1) the sample size and structure for each outcome measure, including numbers of districts, schools, classes, and students; and (2) the nonstandardized unconditional variance components estimated using a three-level model and a two-level model, including estimated variance components at each level of the model, as well as the estimated total variance, all of which are in the original measurement unit of each outcome.

Appendix Table C1 Three-Level vs. Two-Level Model Comparisons: Nonstandardized Unconditional Variance Components

					Nonstandardized Unconditional Variance Components						
					Three-Level Model				Two-Level Model		
Outcomes	Number of districts	Number of schools	Number of classes	Number of students	School	Class	Student	Total	School	Student	Total
Academic Outcomes											
Print Awareness (CLIMBERs)	1	23	47	430	32.2	1.7	70.6	104.4	33.2	71.4	104.6
Blending (CLIMBERs)	1	23	47	430	3.0	0.2	17.0	20.2	3.1	17.1	20.2
Elision (CLIMBERs)	1	23	47	430	0.0	1.1	14.8	15.9	0.5	15.4	15.9
Expressive Vocabulary (CLIMBERs)	1	23	47	430	19.8	32.5	306.2	358.5	38.2	321.1	359.3
Stanford 9 total math scaled score	4	97	200	791	115.1	36.4	1,273.2	1,424.7	131.4	1,293.2	1,424.6
Stanford 9 total reading scaled score	4	97	200	780	108.8	159.0	1,581.9	1,849.6	181.8	1,666.5	1,848.3
Academic-Related Outcomes											
Breakfast participation (adjusted for attendance)	6	100	208	935	193.4	0.0	745.9	939.2	193.5	745.8	939.3
Attendance	6	111	230	831	0.0	0.8	13.0	13.9	0.3	13.5	13.9
Days tardy as a percentage of number of school days enrolled	5	56	112	483	1.1	0.0	13.5	14.6	1.1	13.5	14.6
Stimulus Discrimination: number of trials incorrect	6	111	233	1,141	0.0	0.1	2.6	2.8	0.0	2.7	2.8
Stimulus Discrimination: average trial time	6	111	233	1,141	0.1	0.1	1.3	1.5	0.1	1.4	1.5
Stimulus Discrimination: average viewing time	6	111	233	1,141	0.1	0.1	1.3	1.4	0.1	1.3	1.4
Digit Span: forward and backward combined and scaled by age	6	111	233	1,144	0.2	0.0	9.2	9.4	0.2	9.2	9.4
Verbal Fluency: number of animals named	6	111	233	1,158	1.1	0.9	18.2	20.1	1.4	18.7	20.1
Verbal Fluency: number of things to eat named	6	111	233	1,158	0.8	0.8	17.7	19.3	1.1	18.2	19.3
Verbal Fluency: VF_ani and VF_eat combined	6	111	233	1,158	3.3	2.8	55.1	61.2	4.4	56.7	61.1
Emotional and Behavioral Outcomes	6	111	233	927	0.0	0.0	0.1	0.1	0.0	0.1	0.1
Pediatric Symptom Checklist (PSC) status, 0=non-PSC case 1=PSC case	6	111	233	927	0.6	0.6	27.1	28.3	0.9	27.4	28.3
Sum of 17 PSC questions											
Conners' ADHD Index	6	111	226	1,050	0.9	9.1	107.3	117.3	4.8	112.5	117.3
Cognitive Problems/Inattention	6	111	226	1,073	0.6	4.4	128.8	133.9	2.8	131.2	133.9
Hyperactivity	6	111	225	1,045	0.0	7.6	95.7	103.3	2.0	101.2	103.2
Oppositional Behavior	6	111	225	1,039	0.0	3.8	100.5	104.3	1.8	102.6	104.3
Ability to Focus	6	111	228	1,092	0.0	0.3	1.9	2.2	0.1	2.1	2.2
Ability to Follow Instructions	6	111	228	1,092	0.0	0.3	1.8	2.1	0.1	2.0	2.1
Health Outcomes											
Body Mass Index percentile	6	111	233	1,150	0.0	0.0	784.4	784.4	0.0	784.4	784.4
At risk of overweight	6	111	233	1,150	0.0	0.0	0.2	0.2	0.0	0.2	0.2
Considered overweight	6	111	233	1,150	0.0	0.0	0.1	0.1	0.0	0.1	0.1
Weight status	6	111	233	1,150	0.0	0.0	0.6	0.6	0.0	0.6	0.6
Height	6	111	233	1,151	0.8	0.4	44.9	46.0	0.9	45.1	46.0
Weight	6	111	233	1,151	1.5	1.5	83.1	86.2	2.2	84.0	86.2

Sources: Where indicated, data are from the CLIMBERs database; all other data are from the School Breakfast Pilot Project (SBPP) year 1 follow-up database.

Notes: Estimated values for the variance components 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; all analyses for outcomes from the SBPP database also include indicator variables for school districts in the study sample.

Appendix D
Proofs of the Relationship between Three-Level Models and Two-Level Models in Terms of Precision

This appendix demonstrates the following: (1) that the minimum detectable effect size (MDES) for the two-level model is equivalent to the MDES for the three-level model; (2) when data in an experiment are generated by a three-level model (e.g., students nested within classrooms, classrooms nested within schools with randomization at level 3) but are analyzed using a two-level hierarchical model, the results will be consistent but inefficient; and (3) in special cases the two-level and three-level results will be identical, yielding consistent and asympotically efficient results.

Minimum Detectable Effect Size

We begin by deriving the expected mean squares for the "true three-level model" and then for the "two-level model reflecting a three-level data structure."

Three-Level Model

This model corresponds to Design A discussed in Part IV of the paper. Let Y_{ijk} be the observed outcome for student $i = \{1,..., N_A\}$ in classroom $k = \{1,...,K\}$ in school $j = \{1,...,J\}$. The model is then

$$Y_{ikj} = \pi_{0kj} + e_{ikj} \qquad e_{ikj} \sim N(0, \sigma_A^2) \qquad \text{(Level 1)} \quad \text{(D.1)}$$

$$\pi_{0kj} = \beta_{00j} + r_{0kj} \qquad r_{0kj} \sim N(0, \gamma_A^2) \qquad \text{(Level 2)} \quad \text{(D.2)}$$

$$\beta_{00j} = \gamma_{000} + \gamma_{10j} T_j + u_{00j} \qquad u_{00j} \sim N(0, \tau_A^2) \qquad \text{(Level 3)} \quad \text{(D.3)}$$

where π_{0kj} is the mean for classroom k in school j;
β_{00j} is the mean for school j;
γ_{000} is the grand mean;
T_j = 1 if school j has been assigned to the experimental condition, 0 if control;
e_{ikj} is a deviation associated with each student;
r_{0kj} is a deviation associated with each classroom;
u_{00j} is a deviation associated with each school;
σ_A^2 is the between-student variance within classrooms;
γ_A^2 is the between-classroom variance within schools; and
τ_A^2 is the between-schools variance.

Total Sum of Square (TSS):

$$TSS = \sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}(y_{ikj} - \bar{y}_{\bullet\bullet\bullet})^2$$

$$= SS_{between-schools} + SS_{between-classrooms} + SS_{between-students}$$

$$= N_A K\sum_{j}^{J}(\bar{y}_{\bullet\bullet j} - \bar{y}_{\bullet\bullet\bullet})^2 + N_A\sum_{k}^{K}\sum_{j}^{J}(\bar{y}_{\bullet kj} - \bar{y}_{\bullet\bullet j})^2 + \sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}(y_{ikj} - \bar{y}_{\bullet kj})^2 \quad \text{(D.4)}$$

$$= \left[N_A KJ\left(\tau_A^2 + \frac{\gamma_A^2}{K} + \frac{\sigma_A^2}{N_A K}\right)\left(1 - \frac{1}{J}\right)\right] + \left[N_A K\left(\gamma_A^2 + \frac{\sigma_A^2}{N_A}\right)\left(1 - \frac{1}{K}\right)\right] + \left[N_A KJ\sigma_A^2\left(1 - \frac{1}{N_A}\right)\right]$$

Expected between-schools mean square (level 3):

$$E(MS_{b/schools}) = E(SS_{b/schools}/(J-1))$$

$$= E\left(\frac{N_A K\sum_{j}^{J}(\bar{y}_{\bullet\bullet j} - \bar{y}_{\bullet\bullet\bullet})^2}{J-1}\right) = \frac{N_A KJ\left(\tau_A^2 + \frac{\gamma_A^2}{K} + \frac{\sigma_A^2}{N_A K}\right)\left(1 - \frac{1}{J}\right)}{J-1} \quad \text{(D.5)}$$

$$= \sigma_A^2 + N_A\gamma_A^2 + N_A K\tau_A^2$$

Expected between-classrooms mean square (level 2):

$$E(MS_{b/classrooms}) = E(SS_{b/classoom}/[J(K-1)])$$

$$= E\left(\frac{N_A\sum_{k}^{K}\sum_{j}^{J}(\bar{y}_{\bullet kj} - \bar{y}_{\bullet\bullet j})^2}{J(K-1)}\right) = \frac{N_A KJ\left(\gamma_A^2 + \frac{\sigma_A^2}{N_A}\right)\left(1 - \frac{1}{K}\right)}{J(K-1)} \quad \text{(D.6)}$$

$$= \sigma_A^2 + N_A\gamma_A^2$$

Expected between-students mean square (level 1):

$$
\begin{aligned}
E(MS_{b/Students}) &= E(SS_{b/Students}/[KJ(N_A-1)]) \\
&= E\left(\frac{\sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}(y_{ikj}-\bar{y}_{\bullet kj})^2}{KJ(N_A-1)}\right) = \frac{N_A KJ\sigma_A^2\left(1-\frac{1}{N_A}\right)}{KJ(N_A-1)} \qquad \text{(D.7)} \\
&= \sigma_A^2
\end{aligned}
$$

Two-Level Model Reflecting a Three-Level Data Structure

Let Y_{ab}^* be the observed outcome for student $a = \{1,..., N_B)$ in school $b = \{1,...,J)$. Then

$$Y_{ab}^* = \beta_{0b}^* + e_{ab}^* \qquad e_{ab}^* \sim N(0, \sigma_B^2) \qquad \text{(Level 1)} \qquad \text{(D.8)}$$

$$\beta_{0b}^* = \gamma_{00}^* + \gamma_{01}T_b + r_{0b}^* \qquad r_{0b}^* \sim N(0, \tau_B^2) \qquad \text{(Level 2)} \qquad \text{(D.9)}$$

where $\beta *_{0b}$ is the mean for school *b*;

$\gamma *_{00}$ is the grand mean;

T_b = 1 if school *j* has been assigned to the experimental condition, 0 if control;

e_{ab}^* is a deviation associated with each student;

r_{0b}^* is a deviation associated with each school;

σ_B^2 is the between-student variance within schools; and

τ_B^2 is the between-school variance.

Since the two-level model (with students at level 1 and schools at level 2) is being used to model a three-level data structure (with students nested within classrooms nested within schools), it follows that $N_B = N_A K$ and the total number of schools, J, stays the same for both models.

The total sum of squares is by definition the same as that for the three-level model.

Expected between-school mean square (level 2):

$$E\left(MS^{*}_{between-schools}\right) = E\left(SS^{*}_{between-schools}/(J-1)\right)$$

$$= E\left(\frac{N_B \sum_{b}^{J} (\bar{y}^{*}_{\bullet b} - \bar{y}^{*}_{\bullet\bullet})^2}{J-1}\right)$$

$$= \frac{N_B J(\tau_B^2 + \frac{\sigma_B^2}{N_B})(1-\frac{1}{J})}{J-1}$$

$$= \sigma_B^2 + N_B \tau_B^2$$

(D.10)

Note it is also the case that
$E\left(MS^{*}_{between-schools}\right) = E\left(SS^{*}_{between-schools}/(J-1)\right)$

$$= E\left(\frac{N_A K \sum_{b}^{J} (\bar{y}_{\bullet\bullet b} - \bar{y}_{\bullet\bullet\bullet})^2}{J-1}\right)$$

$$= \frac{N_A K J\left(\tau_A^{\,2} + \frac{\gamma_A^{\,2}}{K} + \frac{\sigma_A^{\,2}}{N_A K}\right)\left(1-\frac{1}{J}\right)}{J-1}$$

$$= \sigma_A^{\,2} + N_A \gamma_A^{\,2} + N_A K \tau_A^{\,2}$$

(D.11)

Expected within-schools mean square (level 1):

$$
\begin{aligned}
E(MS^{*}_{within-schools}) &= E(SS^{*}_{within-schools}/[J(N_B-1)]) \\
&= E\left(\frac{\sum_{a}^{N_B}\sum_{b}^{J}(y^{*}_{ab}-\bar{y}^{*}_{\bullet b})^2}{J(N_B-1)}\right) = E\left(\frac{\sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}(y_{ikj}-\bar{y}_{\bullet\bullet j})^2}{J(N_AK-1)}\right) \\
&= \frac{\sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}E(y_{ikj}-\bar{y}_{\bullet kj}+\bar{y}_{\bullet kj}-\bar{y}_{\bullet\bullet j})^2}{J(N_AK-1)} \\
&= \frac{\sum_{i}^{N_A}\sum_{k}^{K}\sum_{j}^{J}\left[E(y_{ikj}-\bar{y}_{\bullet kj})^2+E(y_{\bullet kj}-\bar{y}_{\bullet\bullet j})^2\right]}{J(N_AK-1)} \\
&= \frac{N_AKJ\left[\sigma_A^{\,2}\left(1-\frac{1}{N_A}\right)+\left(\gamma_A^2+\frac{\sigma_A^{\,2}}{N_A}\right)\left(1-\frac{1}{K}\right)\right]}{J(N_AK-1)} \\
&= \sigma_A^{\,2}+\frac{N_A(K-1)}{N_AK-1}\gamma_A^2
\end{aligned}
$$

(D.12)

Note that

$$\hat{\sigma}_B^{\,2} = E(MS^{2}_{within-schools}) \qquad \text{(D.13)}$$

and Equation D.10 shows that

$$E(MS^{*}_{between-schools}) = \sigma_B^2 + N_B\tau_B^2$$

(D.14)

It follows by substituting Equations D.13 and D.14 into Equations D.12 and D.11 that

$$\sigma_B^{\,2} = \sigma_A^{\,2}+\frac{N_A(K-1)}{N_AK-1}\gamma_A^2$$

(D.15)

and

$$\tau_B^2 = \frac{E(MS^*_{Between}) - E(MS^*_{Within})}{N_B}$$

$$= \frac{\sigma_A^2 + N_A\gamma_A^2 + N_A K\tau_A^2 - \left(\sigma_A^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2\right)}{N_A K}$$

$$= \tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2$$

(D.16)

We are now prepared to show that the MDES for the two-level model is equivalent to the MDES for the three-level model. Recall that Part IV of the paper presented the expressions for the MDES for a three-level model (design A) and a two-level model (design B):

For the three-level model:

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

(D.17)

For the two-level model reflecting the three-level data structure:

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

(D.18)

As shown below, given the same data structure (i.e., $N_B = N_A K$, and same number of schools, J, for both models), the MDES for Design A is equivalent to that for Design B.

Proof:

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

$$= \frac{\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}} * \frac{1}{\sqrt{\tau_A^2 + \gamma_A^2 + \sigma_A^2}}}{\sqrt{\frac{\tau_B^2}{J} + \frac{\sigma_B^2}{JN_B}} * \frac{1}{\sqrt{\tau_B^2 + \sigma_B^2}}}$$

$$= \frac{\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}} * \frac{1}{\sqrt{\tau_A^2 + \gamma_A^2 + \sigma_A^2}}}{\sqrt{\frac{\tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2}{J} + \frac{\sigma_A^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2}{JKN_A}} * \frac{1}{\sqrt{\tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2 + \sigma_A^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2}}}$$

$$= \frac{\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}}}{\sqrt{\frac{\tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2}{J} + \frac{\sigma_A^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2}{JKN_A}}}$$

$$= \frac{\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}}}{\sqrt{\frac{\tau_A^2}{J} + \frac{\gamma_A^2}{JK} + \frac{\sigma_A^2}{JKN_A}}}$$

$$= 1$$

Three-Level Data Analyzed Using a Two-Level Model

In this section we suppose that the data in an experiment are generated by a three-level model (e.g., students nested within classrooms, classrooms nested within schools) with randomization at level 3. However, only student and school level data are available to the analyst. The analyst therefore incorrectly uses a two-level hierarchical model to analyze three-level data. We show that the two-level results are consistent but inefficient.

Model

We can write the "correct" model as

$$Y_{ikj} = \alpha + \theta T_j + \mathbf{C}_{ikj}^T \boldsymbol{\delta} + u_j + r_{kj} + e_{ikj},$$

(D.19)

where

Y_{ikj} is the outcome for student *i* within classroom *k* of school *j*;

T_j = 1 if school *j* has been assigned to the experimental condition, 0 if control;

$\mathbf{C}_{ikj}^T$ is a row vector of known pre-treatment covariates;

$\alpha, \theta, \boldsymbol{\delta}$ are unknown regression parameters to be estimated;

$u_j \sim N(0, \tau_A^2)$ is a school-specific random effect;

$r_{kj} \sim N(0, \gamma_A^2)$ is a classroom-specific random effect; and

$e_{jk} \sim N(0, \sigma_A^2)$ is a child-specific random effect.

The random effects are mutually independent of each other and of the predictors in the model.

Writing equation (D.18) in vector notation, we have:

$$Y_{ikj} = (1 \;\; T_j \;\; \mathbf{C}_{ikj}^T)\begin{pmatrix}\alpha\\ \theta\\ \boldsymbol{\delta}\end{pmatrix} + u_j + r_{kj} + e_{ikj}$$
$$= \mathbf{X}_{ikj}^T\boldsymbol{\beta} + u_j + r_{kj} + e_{ikj},$$
(D.20)

where $\mathbf{X}_{ikj}^T = (1 \;\; T_j \;\; \mathbf{C}_{ikj}^T)$ and $\boldsymbol{\beta}^T = (\alpha \;\; \theta \;\; \boldsymbol{\delta}^T)$. Next, we stack all the outcomes within a single school into a column vector of length n_j, the number of students school j. We therefore have

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{1}_j u_j + \mathbf{A}_j\mathbf{r}_j + \mathbf{e}_j,$$
(D.21)

Where:

$\mathbf{Y}_j$ is a vector of length n_j having elements Y_{ikj};

$\mathbf{X}_j$ is a matrix having rows $\mathbf{X}_{ikj}^T$;

$\mathbf{1}_j$ is a vector of length n_j having elements equal to unity;

$\mathbf{r}_j$ is a column vector having elements r_{kj};

$\mathbf{A}_j = \bigoplus_{k=1}^{K_j}\mathbf{1}_{jk}$ (the operator "$\oplus$" stacks elements along the main diagonal of a matrix);

$\mathbf{1}_{kj}$ is a vector of length n_{kj} having elements equal to unity, n_{kj} being the number of students in classroom *kj*; and

$\mathbf{e}_j$ is a vector of length n_j having elements e_{ikj}.

Based on (D.21) we see that:

$$Var(\mathbf{Y}_j) = Var(\mathbf{1}_j u_j + \mathbf{A}_j\mathbf{r}_j + \mathbf{e}_j) = \tau_A^2\mathbf{1}_j\mathbf{1}_j^T + \gamma_A^2\bigoplus_{k=1}^{K_j}\mathbf{1}_{kj}\mathbf{1}_{kj}^T + \sigma^2\mathbf{I}_j \equiv \mathbf{V}_j,$$
(D.22)

where $\mathbf{I}_j$ is the identity matrix having dimension *j*.

It is well known that, under our assumptions and with $\mathbf{V}_j$ known, the generalized least squares estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{-1} \mathbf{Y}_j$$

(D.23)

is the unique minimum variance unbiased estimator and is efficient, having variance-covariance matrix (Seber and Lee, 2003)

$$Var(\hat{\boldsymbol{\beta}}_{GLS}) = \left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{-1} \mathbf{X}_j \right)^{-1}.$$

(D.24)

The problem with GLS, of course, is that $\mathbf{V}_j$, which depends on $(\tau_A^2, \gamma_A^2, \sigma_A^2)$, is not known. Instead, the conventional practice in multilevel data analysis (Raudenbush and Bryk, 2002, Chapter 14) is to substitute maximum likelihood estimators (MLE) $(\hat{\tau}_A^2, \hat{\gamma}_A^2, \hat{\sigma}_A^2)$ into (D.22), yielding $\hat{\mathbf{V}}_j = \hat{\tau}_A^2 \mathbf{1}_j \mathbf{1}_j^T + \hat{\gamma}_A^2 \bigoplus_{k=1}^{K_j} \mathbf{1}_{kj} \mathbf{1}_{kj}^T + \hat{\sigma}_A^2 \mathbf{I}_j$ and therefore generating the "feasible GLS" or "FGLS" estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = \left(\sum_j \mathbf{X}_j^T \hat{\mathbf{V}}_j^{-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \hat{\mathbf{V}}_j^{-1} \mathbf{Y}_j .$$

(D.25)

As the number of schools increases, the MLE $(\hat{\tau}_A^2, \hat{\gamma}_A^2, \hat{\sigma}_A^2)$ converge to their true values $(\tau_A^2, \gamma_A^2, \sigma_A^2)$ so that $\hat{\mathbf{V}}_j$ converges to $\mathbf{V}_j$ and $\hat{\boldsymbol{\beta}}_{FGLS}$ converges to $\hat{\boldsymbol{\beta}}_{GLS}$. Thus, $\hat{\boldsymbol{\beta}}_{FGLS}$, while not generally unbiased or efficient in small samples, is consistent and asymptotically efficient (Seber and Lee, 2003).

Suppose now that we mis-specify the model by ignoring classroom variation. We believe, falsely, that the model is

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{1}_j u_j^* + \mathbf{e}_j^*, \quad Var(\mathbf{Y}_j) = \tau_B^2 \mathbf{1}_j \mathbf{1}_j^T + \sigma_B^2 \mathbf{I}_j = \mathbf{V}_j^*$$

(D.26)

assuming

$u_j^* \sim N(0, \tau_B^2)$ is a school-specific random effect;

$e_{kj}^* \sim N(0, \sigma_B^2)$ is a child-specific random effect.

The ostensible between-school variance, τ_B^2, based on the mis-specified two-level model, is not equivalent to the actual between-school variance τ_A^2, and the ostensible within-school variance σ_B^2 is not equivalent to σ_A^2 (see above for a simple example). Nor is the ostensible variance of the outcome $\mathbf{V}_j^*$ in general equivalent to the actual variance, $\mathbf{V}_j$.

What happens to inferences about the regression coefficients when one adopts the mis-specified model? In this case, one will use the feasible GLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS*} = \left(\sum_j \mathbf{X}_j^T \hat{\mathbf{V}}_j^{*-1} \mathbf{X}_j \right)^{-1} \sum_k \mathbf{X}_j^T \hat{\mathbf{V}}_j^{*-1} \mathbf{Y}_j \, .$$

(D.27)

As the number of schools increases, this estimator will converge in probability to its GLS counterpart

$$\hat{\boldsymbol{\beta}}_{GLS*} = \left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{Y}_j$$

(D.28)

having mean

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{GLS*}) &= E\left[\left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{Y}_j \right] \\ &= E\left[\left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \left(\mathbf{X}_j \boldsymbol{\beta} + \mathbf{1}_j u_j + \mathbf{A}_j \mathbf{r}_j + \mathbf{e}_j \right) \right] \\ &= \boldsymbol{\beta} + \left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j \right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} E\left(\mathbf{1}_j u_j + \mathbf{A}_j \mathbf{r}_j + \mathbf{e}_j \right) \\ &= \boldsymbol{\beta}. \end{aligned}$$

(D.29)

Thus, the mis-specified model produces a consistent estimator of the regression coefficients. Note that it can also be shown that estimating the model using OLS will yield an unbiased estimate of the regression coefficients.

The asymptotic variance of the mis-specified estimator will be

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_{GLS*}) &= Var\left[\left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j\right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{Y}_j\right] \\
&= Var\left[\left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j\right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \left(\mathbf{X}_j \boldsymbol{\gamma} + \mathbf{1}_j u_j + \mathbf{A}_j \mathbf{r}_j + \mathbf{e}_j\right)\right] \\
&= \left(\sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j\right)^{-1} \sum_j \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{V}_j \mathbf{V}_j^{*-1} \mathbf{X}_j \left(\sum_k \mathbf{X}_j^T \mathbf{V}_j^{*-1} \mathbf{X}_j\right)^{-1}
\end{aligned}
$$

(D.30)

and will therefore not generally be asymptotically efficient.

Special Case with Balanced Data and No Covariates at Levels 1 or 2

In the special case of balanced data and no covariates at levels 1 or 2, it can be shown that the estimators from the three-level model (D.25) and the mis-specified two-level model (D.28) will be identical (and identical to OLS). Moreover, the variance-covariance matrices (D.24 and D.30) will be identical. This implies that in these special cases the two-level model will yield results that are consistent and asymptotically efficient. To see this, revise the level 3 model (D.3) to include a treatment contrast, yielding

$$
\beta_{00j} = \gamma_{000} + \gamma_{001} T_j + u_{00j}, \quad u_{00j} \sim N(0, \tau_A^2) \qquad \text{(Level 3)}
$$

(D.31)

where $T_j = 1/2$ if school j is assigned to the experimental condition and $T_j = -1/2$ if school j is assigned to the control condition, with $J/2$ schools in each condition. Then it is easy to show that

$$Var\begin{pmatrix}\hat{\gamma}_{000}\\ \hat{\gamma}_{001}\end{pmatrix}=\left(\sum_j \mathbf{X}_j^T\mathbf{V}_j^{-1}\mathbf{X}_j\right)^{-1}=\begin{pmatrix}J & 0\\ 0 & J/4\end{pmatrix}^{-1}\begin{pmatrix}\tau_A^2+\frac{\gamma_A^2+\sigma_A^2/N_A}{K}\end{pmatrix}.$$

(D.32)

Now revise the mis-specified level 2 model (D.9) similarly:

$$\beta_{0b}^* = \gamma_{00}^* + \gamma_{01}^* T + r_{0b}^* \qquad r_{0b}^* \sim N(0, \tau_B^2).$$

(D.33)

We will then find that

$$Var\begin{pmatrix}\hat{\gamma}_{00}^*\\ \hat{\gamma}_{01}^*\end{pmatrix}=\left(\sum_j \mathbf{X}_j^T\mathbf{V}_j^{*-1}\mathbf{X}_j\right)^{-1}=\begin{pmatrix}J & 0\\ 0 & J/4\end{pmatrix}^{-1}\begin{pmatrix}\tau_B^2+\sigma_B^2/N_B\end{pmatrix}$$

(D.34)

Using the results regarding the MDES above, we can show that D.32 and D.34 are equal:

Note that

By design, $N_B = N_A K$

D.15 shows that

$$\sigma_B{}^2 = \sigma_A{}^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2$$

And D.16 shows that

$$\tau_B^2 = \tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2$$

Therefore,

$$\tau_B^2 + \sigma_B^2 / N_B = \tau_A^2 + \frac{N_A - 1}{N_A K - 1}\gamma_A^2 + [\sigma_A{}^2 + \frac{N_A(K-1)}{N_A K - 1}\gamma_A^2]/(N_A K)$$

$$= \tau_A^2 + [\frac{N_A - 1}{N_A K - 1} + \frac{N_A(K-1)}{(N_A K - 1)N_A K}]\gamma_A^2 + \sigma_A{}^2/(N_A K)$$

$$= \tau_A^2 + \frac{(N_A - 1)N_A K + N_A(K-1)}{(N_A K - 1)N_A K}\gamma_A^2 + \sigma_A{}^2/(N_A K)$$

$$= \tau_A^2 + \frac{N_A N_A K - N_A}{(N_A K - 1)N_A K}\gamma_A^2 + \sigma_A{}^2/(N_A K)$$

$$= \tau_A^2 + \frac{\gamma_A^2}{K} + \sigma_A{}^2/(N_A K)$$

$$= \tau_A^2 + \frac{\gamma_A^2 + \sigma_A^2 / N_A}{K}$$

Hence, D.32 and D.34 are equal.

When the data are nearly balanced, the estimator based on the mis-specified model will still be consistent (Equation D.29), and we can anticipate that the estimators and their variance-covariance matrices will be approximately equal.

References

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

American Institutes for Research and MDRC. 2006. "Data Collection and Data Analysis Plan for a Professional Development Impact Study." Unpublished paper. Prepared for the Institute of Education Sciences, U.S. Department of Education, Contract No. ED-01-CO-0026/0020.

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

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

Boruch, Robert F. (ed.) 2005. *Place-Based Trials: Experimental Tests of Public Policy*. Thousand Oaks, CA: Sage.

Boruch, Robert F., and Ellen Foley, 2000. "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials." In Leonard Bickman (ed.), *Validity and Social Experimentation: Donald Campbell's Legacy*, Volume 1. Thousand Oaks, CA: Sage Publications.

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.

Conners, C. K. 2000. *Conners' Rating Scales-Revised Technical Manual*. North Tonawanda, NY: Multi-Health Systems.

Detterman, D. 1988. *Cognitive Abilities Tests.* Cleveland, OH: Case Western Reserve University, Department of Psychology.

Donner, Allan, and Neil Klar. 2000. *Design and Analysis of Group Randomization Trials in Health Research*. London: Arnold.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Gardner, W., J. M. Murphy, G. Childs, K. Kelleher, M. Pagano, M. Jellinek, T. K. McInerny, R. C. Wasserman, P. Nutting, and L. Chiapetta, 1999. "The PSC-17: A Brief Pediatric Symptom Checklist with Psychosocial Problems Subscales: A Report from PROS and ASPN." *Ambulatory Child Health* 5, 3: 225-236.

Hedberg, Eric C., Rafael Santana, and Larry V. Hedges. 2004. "The Variance Structure of Academic Achievement in America." Presentation to the 2004 Annual Meeting of the American Educational Research Association.

Hedges, Larry V., and Eric C. Hedberg. 2007. "Intra-Class Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29: 60-87.

Hill, Carolyn J., Howard S. Bloom, Alison Black, and Mark Lipsey. Forthcoming. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*.

Jacoby, E., S. Cueto, and E. Pollitt. 1996. "Benefits of a School Breakfast Programme among Andean Children in Huaraz, Peru." *Food and Nutrition Bulletin* 17, 1: 54-64.

Lonigan, C., R. Wagner, J. Torgesen, and C. Rashotte. 2002. *Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP)*. Tallahassee, FL: Authors.

Murphy, J. M., C. Wehler, M. Pagano, M. Little, R. Kleinman, and M. Jellinek. 1998. "Relationship between Hunger and Psychosocial Functioning in Low-Income American Children." *Journal of the American Academy of Child and Adolescent Psychiatry* 37, 2: 163-170.

Murray, David M. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Murray, David M., and Jonathan L. Blitstein. 2003. "Methods to Reduce the Impact of Intraclass Correlation in Group-Randomized Trials." *Evaluation Review* 27, 1: 79-103.

Murray, David M., and Brian Short. 1995. "Intra-class Correlation among Measures Related to Alcohol Use by Young Adults: Estimates, Correlates and Applications in Intervention Studies." *Journal of Studies on Alcohol* 56, 6: 681-694.

Pollitt, E., N. L. Lewis, C. Garza, and R. Shulman. 1982/83. "Fasting and Cognitive Function." *Journal of Psychiatric Research* 17, 2, 169-174.

Pollitt, E., and R. Mathews. 1998. "Breakfast and Cognition: An Integrative Summary." *American Journal of Clinical Nutrition* 67 (suppl.): 804S-813S.

Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Group Randomized Trials." *Psychological Methods* 2, 2: 173-185.

Raudenbush, S. W., and A. S. Bryk, 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Raudenbush, Stephen W., Andres Martinez, and Jessaca Spybrook. 2007. "Strategies for Improving Precision in Group-Randomized Experiments.**"** *Educational Evaluation and Policy Analysis* 29: 5-29.

Rothbart, M .K., S. A. Ahadi, D. E. Evans, 2000. "Temperament and Personality: Origins and Outcomes." *Journal of Personality and Social Psychology* 78, 1: 122-135.

Schochet, Peter A. 2005. *Statistical Power for Random Assignment Evaluations of Education Programs.* Princeton: Mathematica Policy Research.

Seber, G. A. F., and A. J. Lee, 2003. *Linear Regression Analysis* (Second Edition). New York: John Wiley & Sons.

Siddiqui, Ohidul, Donald Hedeker, Brian R. Flay, and Frank B. Hu. 1996. "Intra-Class Correlation Estimates in a School-Based Smoking Prevention Study: Outcome and Mediating Variables by Gender and Ethnicity." *American Journal of Epidemiology* 144, 4: 425-433.

Simeon, D. T., and S. Grantham-McGregor. 1989. "Effects of Missing Breakfast on the Cognitive Functions of School Children with Differing Nutritional Status." *American Journal of Clinical Nutrition* 49: 646-653.

Ukoumunne, O. C., M. C. Gulliford, S. Chinn, J. A. C. Sterne, and P. F. J. Burney. 1999. "Methods for Evaluating Area-Wide and Organisation-Based Interventions in Health and Health Care: A Systematic Review." *Health Technology Assessment* 3, 5: 1-99.

U.S. Department of Agriculture, Food and Nutrition Service, Office of Analysis, Nutrition, and Evaluation. 2002. *Evaluation of the School Breakfast Program Pilot Project: Findings from the First Year of Implementation,* Special Nutrition Programs Report No. CN-02-SBP.

Visscher, Peter M. 1998. "On the Sampling Variance of Intra-class Correlations and Genetic Correlations." *Genetics* 149: 1605-1614.

Wechsler, D. 1991. *WISC-III Manual*. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.