MDRC Working Papers on Research Methodology

New Empirical Evidence for the Design of Group Randomized Trials in Education

Robin Jacob
University of Michigan

Pei Zhu
Howard S. Bloom
MDRC

December 2009

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

This paper was supported by funding from the W.T. Grant Foundation, the Judith Gueron Fund for Methodological Innovation in Social Policy Research at MDRC, and a staff development grant from Abt Associates, Inc. The authors thank Steve Raudenbush and Brian Jacob for their extremely helpful comments. The ideas presented and positions taken in the paper are solely the responsibility of the authors however, and do not necessarily reflect views of the funders.

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

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

The findings and conclusions in this 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.

Contents

List of Tables and Figures

Table

Figure

Abstract

This paper provides practical guidance for researchers who are designing studies that randomize groups to measure the impacts of educational interventions. The paper (1) provides new empirical information about the values of parameters that influence the precision of impact estimates (intra-class correlations and R-squared values) and includes outcomes other than standardized test scores and data with a three-level structure rather than a two-level structure, and (2) discusses the error (both generalizability and estimation error) that exists in estimates of key design parameters and the implications this error has for design decisions. Data for the paper come primarily 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 four-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 from 6 school districts. Student achievement data from the Reading First Impact Study is also used to supplement the discussion.

Introduction

Group randomized trials have become widely used as an important way to measure the effectiveness of a variety of educational interventions. In such trials, groups of individuals, such as classrooms or schools, rather than individual students, are randomly assigned to treatment or control conditions. Since many educational interventions are intended to change school environments (for example, whole school reform efforts) or classrooms contexts (for example, specific curricular programs, teacher professional development programs), group randomized designs provide an effective way to measure the causal impacts of the interventions. In recent years, the United States Department of Education's Institute for Education Sciences has encouraged the use of randomized trials in its grants funding program and has funded a series of large-scale group randomized studies (for example, see Garet et al., 2008).

Although such studies have the potential to provide important information about the causal impacts of educational interventions, they are also expensive to conduct. It is therefore incumbent upon researchers to design group randomized studies carefully, so that they yield useful information. Important design considerations include (1) the total number of groups to randomize, (2) the average number of individuals to observe per group, (3) the proportion of groups to allocate to treatment and control status, (4) what variables, if any, to use for covariate adjustments, and (5) the categories, if any, by which to block groups before they are randomized. Further design decisions are required for any given study based on the study's specific goals and context. Unfortunately, in practice, researchers often do a poor job of taking design considerations into account. A recent study showed that many government-sponsored studies of educational interventions do not have adequate power because some of the design factors were not properly considered (Spybrook, 2007). Similarly, Hedges (2004) suggests that many studies overestimate the statistical precision of their impact estimates because they fail to take clustering into account.

Over the past several years, a number of papers have been published to provide guidance to researchers designing studies that involve groups (for example, Bloom, 2005; Schochet, 2007; Murray and Blitstein, 2004; and Raudenbush, 1997). 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; and Hedges and Hedberg, 2007). These papers provide the statistical framework for understanding group randomized studies and discuss the implications of the framework for design and analysis.

As these papers describe, group randomization has a multilevel variance and covariance structure, with individual subjects clustered in randomized groups. Many studies in education have a two-level structure and measure impacts on students (Level 1) by randomizing schools (Level 2). The Level 1 variance represents how an outcome varies across individual students within randomized schools. The Level 2 variance represents how the mean value of the individual outcome varies across randomized groups. The Level 1 variance can be designated as σ^2 and the Level 2 variance can be designated as τ^2. Using this framework, the total individual variance across all subjects in all randomized groups equals $\tau^2+\sigma^2$. This information is often expressed as the relationship between the two variances, referred to as an intra-class correlation (ICC), ρ, (Fisher, 1925):

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{1}$$

The intra-class correlation is thus the proportion of total individual subject-level variance that is between randomized groups. As discussed in more detail in this paper, it is also possible to have a three-level variance and covariance structure, such as students nested within classrooms (or by teacher), and classrooms (or teachers) nested within schools. To design group randomized studies that can attain desired levels of precision requires information about the variance at each level.

It is often possible to increase markedly the precision of group randomized studies by adjusting for baseline covariates (see, for example, Bloom, Richburg-Hayes, and Black, 2007, and Murray and Blitstein, 2004). Therefore, in addition to knowledge about the variances, knowledge about 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.

In the health and prevention sciences, information about the values of intra-class correlations, and to a lesser extent, R-squared values at each level, has been catalogued by researchers (for example, Murray and Short 1995; Murray and Blitstein, 2004; and Siddiqui et al., 1996). A repository of this information is maintained by David Murray and his associates.[1] A few researchers have attempted to compile similar information for educational and child development outcomes, but most such information is limited to outcome measures based on standardized achievement test scores from group-administered exams of students in Kindergarten through the twelfth grade (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 (or by teacher).

This paper attempts to broaden the empirical foundation for designing group randomized studies in education. Using two data sets derived from group randomized studies, it

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

builds on previous work by (1) providing estimates of intra-class correlations and R-squared values for outcomes other than standardized test scores, including academic-related outcomes, behavioral outcomes, and health outcomes, and (2) providing estimates of these same parameters based on data using a three-level structure rather than a two-level structure.

The paper also discusses the error (both generalizability error and estimation error) that exists in estimates of key design parameters and the implications this error has for design decisions.

The first section of the paper describes the data sources and defines the outcome measures that are used. The next section presents estimates of intra-class correlations and R-squared values at each level for a series of academic and child outcome measures from these two data sets and 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. Finally, the paper discusses the sources of error in the parameters used to estimate the precision of impacts, 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.

Data Sources, Student Samples, and Outcome Measures

Data for the present paper were obtained from two studies that randomized schools to measure intervention effects on children — the Chicago Literacy Initiative: Making Better Early Readers Study (CLIMBERS) and the School Breakfast Pilot Project (SBPP) — and one study that used a regression discontinuity design, the Reading First Impact Study (RFIS). This section describes the studies, their samples, and the outcome measures used for this 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 School preschool classrooms that serve four-year-old children.[2] These preschool programs were generally associated with elementary schools in the Chicago Public School system. Schools were recruited to participate in the study if they were low performing and had few other early literacy initiatives.

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

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.[3] The goal of the project was to measure the impact of Breakthrough to Literacy at scale on student pre-literacy skills.

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

One goal of this paper is to examine three-level data structures, for students clustered within classrooms clustered within schools. Therefore, only schools with two or more classrooms in the study are included. This limited the analysis sample to 430 preschool students from 47 classrooms in 23 schools.

School Breakfast Pilot Project

This three-year demonstration project (2000-2003) was based on an experimental design that randomized schools within six school districts 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.[4] The goal of the project was to measure the added value of universal free school breakfasts.[5]

Six school districts were chosen for the project from among 136 that applied to participate. The resulting project sample included students in grades 2 through 6 from 138 elementary schools. Within each treatment 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. Findings reported are based on data for 1,151 third-graders from 233 classrooms in 111 schools located in 6 school districts.[6] Outcome measures were obtained from several sources and include academic outcomes, other school-related outcomes, emotional and behavioral outcomes, and health outcomes. These data were used to estimate intra-class correlations for

[3] 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.

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

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

[6] The numbers of students, classrooms, and schools vary by outcomes because of item nonresponse.

the three-level variance structures and corresponding R-squared values from the use of covariates (described later).

Reading First Impact Study

The Reading First Impact Study was a three year (2004-2007) congressionally mandated evaluation of the federal government's Reading First initiative conducted by Abt Associates Inc., MDRC, and their research partners, and sponsored by the United States Department of Education. Reading First, part of the No Child Left Behind Act, provides guidance and funding to low-performing schools and promotes instructional practices that have been validated by scientific research, with the goal of helping all children read at or above grade level by the end of the third grade. The study used a regression discontinuity design that capitalized on the systematic process used by a number of school districts to allocate their Reading First funds to an evaluation of the program's impact on teacher practices and study achievement. Chosen from a pool of eligible sites to participate in the study were 17 districts plus one state program. The final study sample included 248 schools from these 18 sites. Survey, classroom observation, and student achievement data were collected from all 248 schools over the three-year period of the study. This study uses Stanford Achievement Test, Version 10 (SAT 10) reading achievement data that were collected from all first-, second-, and third-grade students in these schools in the fall of 2004, the spring of 2005, the spring of 2006, and the spring of 2007. The data for this paper are limited to 15 sites (14 districts and one state) and 225 schools for which we were able to estimate variance components at the school level.

Measures

Unlike in other studies that focus primarily on group-administered standardized test scores, the outcome measures for the present paper fall into four different categories: (1) academic outcomes (standardized test scores, both group and individually administered), (2) other academic-related outcomes (for example, attendance), (3) student behavior, and (4) health outcomes. These measures are described below.

Academic Outcomes

Four measures of pre-literacy skills were obtained from data for CLIMBERS, based on student scores from the Preschool Comprehensive Test of Phonological and Print Processing (Lonigan, Wagner, Torgesen, and Roshotte, 2002).[7] This test is an individually administered assessment that measures phonological skills that have been shown to be important precursors to reading proficiency. The four measures used in this paper include

[7] 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.

- Print Awareness: The print awareness subtest measures beginning knowledge about written language; for example, the ability to tell what print looks like and how it works. Items measure whether children recognize individual letters, know what sounds letters make, and can 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 a portion of the word. (For example: Say cup without saying /K/.)

- Blending: The blending subtest measures a child's the 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 speaking or writing.

Two measures of third-grade academic performance were obtained from data for the School Breakfast Pilot Project. These measures come from the Stanford Achievement Test, Version 9 (SAT 9).

- Stanford 9 Total Math Scale Score (total test scores in scaled score points).

- Stanford 9 Total Reading Scale Score (total test scores in scaled score points).

One measure of academic performance was used from the Reading First Impact Study. The measure came from the SAT 10:

- Stanford 10 Reading Comprehension Scale Score (total test scores in scaled score points).

Other Academic-Related Outcomes

Other academic-related outcomes include children's behavior and cognitive skills that are related to, or are precursors of, academic achievement. The School Breakfast Pilot Project collected several such measures. The following are included in this paper:

- Attendance: The number of days a child was present at school divided by the total number of school days the child was enrolled.

- Tardiness: The number of days a child was tardy as a percentage of the number of school days enrolled.[8]

- Breakfast Participation (adjusted for attendance): The frequency with which a child attended the school breakfast program, controlling for the child's overall school attendance.

- Stimulus Discrimination: A measure of cognitive performance that assesses a child's ability to distinguish between similar stimuli presented on a computer screen (Detterman, 1988). Three variables are used in these analyses: (1) number of trials incorrect, (2) average viewing time (the total time of viewing stimuli, averaged across all trials), and (3) average trial time (the total viewing and response time, averaged across all trials).

- Digit Span: The digit span subtest of the Wechsler Intelligence Scales for Children III measures cognitive performance and assesses short-term auditory memory and focusing abilities (Wechsler, 1991). Through headphones, a child hears a recorded series of digits. The child then repeats the series back to the tester, forwards in the first part of the task and backwards in the second part of the task. Outcomes are presented in terms of the total number of forward and backward tasks completed correctly, scaled by age.

Three outcome measures are taken from a test of verbal fluency in which children were asked to name as many items as possible in two semantic categories ("animals" and "things to eat"), in a period of 60 seconds for each category. The number of items in a particular category that students name in a given period of time is intended to measure neuropsychological functioning in the areas of long-term verbal memory and retrieval (Simeon and Grantham-McGregor, 1989). Three outcome measures are used in the present paper: (1) verbal fluency (number of animals named), (2) verbal fluency (number of things to eat named), (3) verbal fluency (number of animals named and things to eat named, combined).

Emotional and Behavioral Outcomes

The School Breakfast Pilot Project also provides a wide range of psychosocial and behavioral measures for young children.

Two outcome measures used are from the Pediatric Symptom Checklist (PSC) that was included as part of the Parent Questionnaire in the School Breakfast Pilot Project study (Murphy et al., 1998a). The Pediatric Symptom Checklist was developed as a screening tool

[8] 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.

for psychosocial problems. Outcomes include: (1) A PSC status of 1 if a child is considered psychosocially impaired and a PSC status of 0 if not, and (2) a total PSC score, representing the sum of parents' responses to 17 questions.

Four outcome measures were taken from the Revised Connors Teacher Rating Scale (CTRS-R), part of the The Conners' Rating Scales, which are used to assess psychopathology and behavioral 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 asks teachers to rate children on a variety of behaviors. The outcomes used for this paper are

- Conners' ADHD Index: Identifies children as being at risk for ADHD (Conners, 1997);
- Cognitive Problems/Inattention: High scorers may have more academic difficulties than most individuals their age, problems organizing their work, and difficulty completing tasks or schoolwork, and may 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 always to 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 of the same age.

Two outcomes were taken from the Effortful Control Scale, a subset of questions from the Children's Behavior Questionnaire, a highly differentiated assessment designed to measure temperament in children (Rothbart, 2002). Two subscales are used in this analysis: (1) ability to focus, and (2) ability to follow instructions.

Health Outcomes

Finally, the School Breakfast Pilot Program collected a series of measures of students' health status. Measures used for this paper include

- Body Mass Index Percentile: a direct measure of a child's body mass index.
- At Risk of Overweight: whether or not a child was at risk of being overweight.
- Height (measured in inches).

- Weight (measured in pounds).

New Information About Intra-Class Correlations and R-Squared Values

One of the primary goals of this paper is to provide new information about intra-class correlations and R-squared values at each level for outcomes other than standardized test scores and for data with a three-level variance structure. We begin by reviewing how data on intra-class correlations and R-squared values can be used to design group randomized studies and then present estimates of these parameters from data for the two studies described previously. We then illustrate 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. The present 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 minimum detectable effect size 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. The paper follows this convention.

To choose a minimum detectable effect size for a given study requires an understanding of the effect size's specific context. 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, to 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 only detect much larger impacts, which might be impossible to attain. Hence, the smaller sample would be underpowered statistically. Hill, Bloom, Black, and Lipsey (2008) provide a series of empirical benchmarks for helping to determine an appropriate minimum detectable effect size for educational interventions. There is little such empirical guidance for other fields of intervention research, however.

A minimum detectable effect size is defined in terms of the underlying population's standard deviation for a given outcome measure. For example, a minimum detectable effect size 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 minimum detectable effect size 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

$$MDES = M * \sqrt{Var(impact)} / \sigma_{total} \tag{2}$$

where:

M = a multiplier that depends on the assumed power, significance level, and one- or two-tailed nature of the statistical test, plus the number of degrees of freedom of the study design;

Var (impact) = the variance of the impact estimate;

σ_{total} = the standard deviation of the outcome measure across all individual subjects in the target population (or sample).[9]

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 they do under individual randomization.

Consequently, variance expressions for a group randomized design must account for each variance component. For example, the minimum detectable effect size 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:

$$MDES = M_{(J-2)} * \frac{1}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau^2}{J} + \frac{\gamma^2}{J*K} + \frac{\sigma^2}{J*K*N}} * \frac{1}{\sqrt{\tau^2 + \gamma^2 + \sigma^2}} \tag{3}$$

where:

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

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

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

[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 A for detailed definition relating to M.

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

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

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

K = the harmonic mean number of classrooms per school;

N = the harmonic mean number of students per classroom.[10]

Equation 3 corresponds to Equation 2 in that:

1) ($\tau^2 + \gamma^2 + \sigma^2$) equals the total variance of the outcome measure across all students from all classrooms in all schools, or σ^2_{total}.

2) $\frac{1}{P(1-P)} * (\frac{\tau^2}{J} + \frac{\gamma^2}{J*K} + \frac{\sigma^2}{J*K*N})$ is the variance of the estimated impact and represents the influence of the school-level, classroom-level and student-level variance components and the proportion of groups 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 minimum detectable effect size is

$$MDES = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\tau^2(1-R^2_{sc})}{J} + \frac{\gamma^2(1-R^2_{cl})}{J*K} + \frac{\sigma^2(1-R^2_{st})}{J*K*N}} * \frac{1}{\sqrt{\tau^2 + \gamma^2 + \sigma^2}} \qquad (4)$$

where:

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

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

[10] The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. For example, the harmonic mean of 5 and 7 is 2/(1/5 +1/7)=5.83.

R_{st}^2 = the explanatory power of all covariates for outcome differences across students within classrooms;

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,

$$R_L^2 = \frac{\sigma_{U,L}^2 - \sigma_{C,L}^2}{\sigma_{U,L}^2} \tag{5}$$

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 4 reduces to Equation 3. 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. For example, after controlling for students' socioeconomic status at the student level, it may be the case that there is greater variation among schools in their mathematics achievement scores. Failing to control for socioeconomic status simply masks the existing variation among schools. This increase in unexplained variation at Level 2 or Level 3 would be reflected by a negative value for the relevant R-squared. Specifically, if a covariate is included at Level 1 (but at no other level), R_{st}^2 will be greater than or equal to zero, but R_{cl}^2 and/or R_{sc}^2 could be less than zero.

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 minimum detectable effect size for a three-level variance structure is

$$MDES = \frac{M_{(J-2-C)}}{\sqrt{P(1-P)}} * \sqrt{\frac{\rho_{sc}(1-R_{sc}^2)}{J} + \frac{\rho_{cl}(1-R_{cl}^2)}{J*K} + \frac{(1-\rho_{sc}-\rho_{cl})(1-R_{st}^2)}{J*K*N}} \qquad (6)$$

where:

All parameters are defined as before.

Equation 6 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 unconditional intra-class correlations and the school-level, classroom-level and student-level R-squared values.

Estimation Models

Values for the preceding parameters were estimated from data for CLIMBERS and the School Breakfast Pilot Program (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:[11]

Level 1

$$Y_{ijk} = \pi_{0jk} + \sum_{s>0} \pi_{sjk} X_{sijk} + \varepsilon_{ijk} \qquad (7)$$

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;

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

$$\pi_{0jk} = \beta_{0k} + \gamma_{jk} \qquad (8)$$

where:

[11] All models were estimated by restricted maximum likelihood estimation, using the PROC MIXED procedure in Statistical Analysis Software (SAS).

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

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

Level 3

$$\beta_{0k} = \theta_0 + \theta_1 T_k + (\sum_{m>1} \theta_m Z_{mk}) + \mu_k \tag{9}$$

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 mth school-level covariate for school k;

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

Ideally we would calculate intra-class correlations in the absence of any particular intervention; however, the most complete data on the widest range of measures is available for postintervention outcomes. Using postintervention data also allows us to explore the influence of covariates such as pretest scores on the intra-class correlations and minimum detectable effect sizes. We therefore include an indicator variable for treatment or control status (T_k) in the model, which removes all existing differences between the treatment and control groups (treatment effects) when estimating variance components. In addition, for the School Breakfast Pilot Program, the model removes all differences among the six participating school districts by including indicator variables for them as school-level covariates (Z_{mk}). 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 the level's 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 CLIMBERS preschool sample and the School Breakfast Pilot Project third-grade sample. Findings for other outcomes are from the School Breakfast Pilot Project third-grade sample.

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 ten outcome measures in this category, three have estimated intra-class correlations that equal zero for classrooms and two have estimated intra-class correlations that equal zero for schools. Values for the remaining measures at both the classroom level- and the school level are typically less than 0.05. These values are generally lower than the values presented for the academic level. [12]

[12] Both tardiness and attendance are count variables that are likely to have skewed distributions because they either include a large number of zeros (many students are never tardy) or ones (many students attend

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 may be because, with the exception of the PSC questionnaire, these measures were constructed based on teacher ratings.

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.

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 similar 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 outcomes that exist both within and across school districts. Based on evidence from past empirical studies and new evidence from three evaluation studies, Schochet (2005) concludes 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.

The values in Table 1 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. It can be shown that 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 data because in a two-level model that omits the classroom level some of the classroom-level variance is absorbed by the school level and the student level (for example, see Moerbeek, 2004). Second, the samples of schools for CLIMBERS and the School Breakfast Pilot Project may be more homogenous than those for entire school districts or nationally representative samples which have been used for most related prior research (for example, Hedges and Hedberg, 2007, and Bloom, Richburg-Hayes, and Black, 2007). This hypothesis is consistent with Hedges and Hedberg (2007), who find that the average unadjusted intra-class correlation is lower among low-achieving schools than among

school every day). Estimating variances components using different distributional assumptions may yield different variance estimates than the ones presented here. Presenting variance components for different distributional assumptions is beyond the scope of the current paper, but researchers using attendance and tardiness as outcome measures should take this into consideration when they design the study.

a nationally representative sample of schools, and with Schochet (2005), who finds lower intra-class correlations once district effects have been accounted for.

While little has been reported about intra-class correlations for emotional, behavioral, and health-related intra-class correlations among children in schools, there is 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, schools (for example, Murray and Blitstein, 2004; Ukoumunne et al., 1999; Siddiqui, et al. 1996; and Murray and Short, 1995). The intra-class correlations for these groups and outcomes are much smaller than those for measures of student achievement in schools and range from about 0.01 to 0.05. This is consistent with what we report here.

Overall, Table 1 suggests that school-level intra-class correlations are generally larger than classroom-level intra-class correlations with the exception of the emotional and behavioral outcomes that were based on teacher ratings. Similarly, the findings suggest that the intra-class correlations of traditional academic outcomes are generally larger than those for academic-related outcomes, emotional and behavioral outcomes, and health outcomes.

An area that has received increased attention in the literature in recent years is the issue of scaling, or using test metrics, and the influence that particular metrics can have on outcomes. A depression scale, for example, might be reported as the fraction of items answered correctly, or it could be reported as an Item Response Theory (IRT) equated measure — a measure derived from a model that identifies the probability of a correct response on each item based on the characteristics of both the individual answering the item and the item itself. While the two measures are often highly correlated, they do not always lead to the same results, particularly when one metric is a type of cut-off score (for example, the percentage of students scoring at or above grade level) while the other is a more continuous measure of performance (such as a scaled score on an achievement test). Reardon (2007) has shown, for example, that race and gender gaps can vary quite substantially depending on the metric used to report them. One question is whether the choice of metric also has an influence on the estimated intra-class correlations obtained from the data. For example, would different estimated intra-class correlations result from looking at scaled scores versus percentile ranks?

To explore this question we used Stanford Achievement Test, Version 10 (SAT 10) reading comprehension scores from the Reading First Impact Study. The SAT 10 reports scores in a variety of metrics, including raw scores, percentile rank, normal curve equivalence, stanine scores, and the percentage of students scoring at or above grade level. We calculated unconditional intra-class correlations in each of the metrics for all three years and all three grades of the Reading First Impact Study data. The results are reported in Appendix

B. Across all metrics the intra-class correlation remains relatively stable, with only the percentage of students scoring at or above grade level showing a slightly lower intra-class correlation than the other metrics. Results were similar for each grade and year for which Reading First Impact Study data was available. This suggests that choice of metric has relatively little effect on the estimate of intra-class correlations.

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 post-test, 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 School Breakfast Pilot Project 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 at each level. 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 School Breakfast Pilot Project study 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 School Breakfast Pilot Project study (which used student-level pretests and demographic information as covariates) for Stanford 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) as well as at the school level (0.494 and 0.840).

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 only R-squared values for student level demographic covariates, because pretest data were not available for these outcomes. Adding students' age, gender, ethnicity, and free or reduced lunch status reduced the student-level variance by very little. 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, as demonstrated below, the school-level variance component is usually the primary factor that determines the sample size that is required.

A number of the R-squared values reported in Table 1 are negative. This could be caused by 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, because the omission of the Level 1 covariate was masking variation at Level 2, which would lead to a negative value of the R-squared.

Finally, 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 only occurs 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.

To summarize: For the academic outcomes, including pretest scores and demographic characteristics at the student level provided considerable explanatory power at the student, classroom, and school levels. Including a pretest and demographics at the school level produced R-squared values that ranged from -0.01 to 1.0 at the school level. For the other outcomes, including demographic characteristics only at the student level provided little explanatory power at the student and classroom levels, but for some outcomes the inclusion of demographic characteristics only at the student level explained a substantial proportion of the variance at the school level.

In addition to estimating R-squared values using the set of covariates with the most explanatory power 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 do 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 the R-squared values for those outcomes are smaller than the R-squared values 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 minimum detectable effect sizes for alternative sample designs. For example, how much benefit is gained by adding additional schools to the sample? How much is gained by adding more students within schools to the sample? What happens if we increase the number of classrooms while holding the number of students the same? 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 minimum detectable effect size of the sample used for the analyses presented in this paper. This sample includes 430 students from 47 classrooms in 23 schools in one school district for the CLIMBERS data, which represents about 9 students per classroom and 2 classrooms per school. For this sample and the estimated intra-class correlations and R-squared values reported in Table 1, minimum detectable effect sizes range from about 0.37 to 0.52 standard deviations for the four CLIMBERS outcome measures.

For the School Breakfast Pilot Project 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 minimum detectable effect sizes for its outcome measures range 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. (Findings in the table assume that half of the schools are randomized to treatment status and half are randomized to control status.) 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. The increase in power obtained by increasing the number of schools is substantial. For print awareness, for example, increasing the number of schools from 20 to 100 reduces the minimum detectable effect size from 0.567 standard deviations to 0.254 standard deviations.

Columns two and five show what happens if we hold the number of classroom and schools constant while increasing the number of students per school by increasing the number of students per classroom fivefold (from 5 to 25), resulting in a fivefold increase in the number of students per school. For print awareness this implies a relatively small reduction in the minimum detectable effect size 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, in the number of randomized groups) improves precision by far more than does the same proportional increase in the number of students per school. (See, for example, Bloom, Richburg-Hayes, and Black, 2007).

We can also explore how, given a fixed total number of schools, changing the number of classrooms and students per school influences precision. 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 from 2 to 4 (and thereby doubling the number of students per school from 10 to 20) reduces the minimum detectable effect size for print awareness from 0.567 standard deviations to 0.512 standard deviations. Note that, in this data, simply increasing the number of classrooms per school while holding the number of total students in the school the same would not have any appreciable effect on power because the variance at the classroom level is relatively small. However, for data that have larger classroom-level variance components, increasing the number of classrooms per school while holding the number of students the same could lead to a lower minimum detectable effect size.

By making comparisons such as the ones described in this section, it is possible to begin to assess the relative precision of alternative sample designs for specific outcome measures — and, in this way, to develop and defend a proposed research design.

Accounting for Uncertainty About Intra-Class Correlations

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

However, there are a number of sources of error in the estimates of these parameters that should be considered when making decisions about study design, and these sources of error are often overlooked. The first is generalizability error. Researchers must consider how similar the planned study sample will be to the sample used to estimate the intra-class correlations and R-squared values. 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 will those for different outcome measures.

Another important, and often overlooked, consideration when assessing the appropriateness of estimated intra-class correlations and R-squared values for planning a study is estimation error — the statistical uncertainty that exists about the estimates. As described in more detail below, for estimated intra-class correlations, this uncertainty depends on the number of groups and the number of subjects per group in the estimation sample. In addition, it depends on the true value of the intra-class correlation for the population of interest. A similar problem of uncertainty arises when using estimated values of R-squared to plan a group randomized study. Although a detailed exploration of the uncertainty associated with R-squared estimates is beyond the scope of this paper, a brief discussion of the literature on this topic follows.

Taking this uncertainty into account is especially important when a researcher might otherwise have confidence in an estimated intra-class correlation or R-squared value because it comes from the same population and it 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 inherent in intra-class correlations for two-level research designs. Two-level designs are considered because most studies have employed them and because the statistical properties of their intra-class correlations are relatively well understood. The discussion of uncertainty proceeds as follows: (1) It first describes how standard errors and confidence intervals can be calculated for estimates of two-level intra-class correlations; (2) it then examines the factors that influence these indicators of uncertainty; (3) it illustrates their implications for the findings from the CLIMBERS and School Breakfast Pilot Project studies; and (4) it decomposes the uncertainty in intra-class correlations into generalizability error and estimation error.

Estimating Uncertainty for Intra-Class Correlations

According to Siddiqui et al. (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:[13]

$$Var(\hat{\rho}) = \frac{2(1-\hat{\rho})^2[1+(N-1)\hat{\rho}]^2}{N(N-1)J} \tag{10}$$

where:

$\hat{\rho}$ = the estimated intra-class correlation;

N = the harmonic mean number of individuals per group;

J = the total number of groups.

The standard error of the estimated intra-class correlation equals the square root of the expression in Equation 10. 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. As noted above, even the most precise results from a sample of white students may not generalize to a sample of black students because the true intra-class correlations between the groups can be different. While the estimates presented here cannot take this variation into account, later we use data from the Reading First Impact Study, which were collected from 225 schools from 15 different sites (14 districts and one state), to provide some empirical evidence about the potential magnitude of generalizability error.

Table 4 illustrates how the standard error derived from Equation 10 varies with $\hat{\rho}$, N, and J. First, as the number of groups (J) increases, the standard error of the estimated intra-class correlation decreases. In fact, Equation 10 implies that 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 group, the estimated standard error of the intra-class correlation decreases from 0.130 to 0.058 (by the square root of five) as the number of groups quintuples from 10 to 50.

Second, as the number of individuals per group (N) increases, the standard error of the estimated intra-class correlation also decreases, although this relationship is more

[13] Equation 10 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.

complex than that for the number of groups. For example, with an intra-class correlation of 0.5 and a total of 10 groups, the estimated standard error of the intra-class correlation decreases from 0.130 to 0.115 as the number of individuals per group quintuples from 10 to 50.

These results illustrate that a proportional increase in the number of groups reduces the standard error of the intra-class correlation by far more than does the same proportional increase in the number of individuals per group. Hence, the relative influence of groups 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.

Finally, 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 groups and 10 individuals per group, 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 group-level variance component, τ^2.

For example, assume that an intra-class correlation was estimated from a sample of 50 groups with 10 individuals per group, using a two-level model with no covariates. If the estimated intra-class correlation (the point estimate) were 0.20, then, according to Table 4, 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 groups minus one (J-1). This implies 49 degrees of freedom for the present example. For a t distribution with 49 degrees of freedom the correct multiple is 2.01, thus 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.

Uncertainty About Intra-Class Correlations for the Present Paper

Table 5 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 School Breakfast Pilot Project study. (Equation 10 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 0.418; and 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.

In comparison, intra-class correlations from the School Breakfast Pilot Project study 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 that for estimates from the CLIMBERS sample. For participation in the school breakfast program, the School Breakfast Pilot Project 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

One way to account for the uncertainty inherent in estimations of intra-class correlations is to assess sample size requirements using not only the point estimate of the intra-class correlation (as is usually done in practice) but also the upper and lower bound of the confidence interval. Although the best single estimate of the sample size requirement is that based on the point estimate for the intra-class correlation, depending on the uncertainty that exists about an estimate, it may 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.

Table 6 illustrates the implications of uncertainty for designing a group randomized study in the CLIMBERS and School Breakfast Pilot Project data. The first column in the table lists the predicted minimum detectable effect size for an illustrative research design given the lower bound of the confidence interval of the intra-class correlation for each outcome measure in Table 5. 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 co-variates for each outcome measure (those used for Tables 1 and 3).

Note that the width of confidence intervals for minimum detectable effect sizes 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 confidence interval of minimum detectable effect sizes for blending (from CLIMBERS) is quite wide, ranging from 0.230 to 0.329 standard deviations. In contrast, the confidence interval of minimum detectable effect sizes for school breakfast participation (from the SBPP study), is much narrower, ranging from 0.275 to 0.317 standard deviation.

Table 7 moves the discussion of uncertainty a step further by translating the findings in Table 6 into their implications for the number of randomized schools needed to achieve a minimum detectable effect size 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.

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 School Breakfast Pilot Project 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 minimum detectable effect size of 0.25 standard deviations.

Two main factors create the preceding differences in uncertainty about required sample sizes. First, the CLIMBERS sample has fewer schools from which to estimate an intra-class correlation than does the School Breakfast Pilot Project 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.

This part of the paper has considered how to quantify the uncertainty that exists about intra-class correlations as a result of 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 requires that researchers 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. It is rare for researchers to report estimates of the precision of R-squared values because the distributions depend on unknown parameters (Press and Zellner, 1978; Ohtani, 1999). Although a full exploration of this topic is beyond the scope of this paper, methodological work in this area has been done. For example, Helland (1987) proposes a simple method for approximating a confidence interval for an R-squared value. Cardouss and Giles (1992) derive the exact distribution of R-squared values in regression models where the error is autocorrelated. Using Monte Carlo simulations, Ohtani (1999) has shown that accurate estimates of the standard error of an R-squared value can be obtained via bootstrap methods (which construct resamples of the observed dataset by random sampling with replacement from the original dataset).

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.

Generalizability Error

Up to this point, we have been considering only the uncertainty that exists in estimates of intra-class correlations that can be attributed to statistical estimation error. But as already noted, as much, or more, of the variability in estimated intra-class correlations may be ascribable to true differences among the samples — in other words to generalizability error. To get a sense of the magnitude of this error in estimates of intra-class correlations we again made use of data collected as a part of the Reading First Impact Study, for which we have data available from 225 schools in 15 different sites (14 districts and one state). We use this data to explore how much of the variability in the estimates of the intra-class correlations across the 15 sites was due to sampling or estimation error and how much could be attributed to true variation in the intra-class correlations across the different sites in the sample (generalizability error).

Table 8 shows the full sample intra-class correlation and the estimated intra-class correlations (along with associated standard errors and 95 percent confidence intervals) from each of the 15 sites in the sample for first-graders in 2005. The table indicates that the estimated intra-class correlations vary considerably from site to site, with a low of 0.018 for Site 4 and a high of 0.208 for Site 12. It is apparent that at least some of variability in the estimated intra-class correlations across sites is due to estimation error since the standard error of the intra-class correlation estimates varies considerably across site. The estimation error is in part due to the variation in the number of schools per site — Site 14 has only 6 schools, for example, while Site 6 has 29. However, part of the variability across sites is almost certainly caused by true differences in the population of students and the characteristics of the schools across these sites. In other words, at least some of the variation in the estimates is due to true differences in the underlying intra-class correlations in these sites.

To get a sense of how much of the variability can be attributed to estimation error, and how much to generalizability error, we used hierarchical linear modeling (Raudenbush and Bryk, 2002). We estimated unconditional two-level models (with the 225 schools nested within the 15 sites) and specified that the Level 1 variance was known and equal to the square of standard errors of the intra-class correlations shown in Table 8. Hierarchical linear modeling then provided estimates of the total variance at Level 2 (that is, generalizability error or the true variability across the 15 sites) as well as the proportion of total error that was due to variance across sites (what Raudenbush and Bryk refer to as reliability). The total error was obtained by dividing the estimated generalizability error by the estimated reliability. Estimation error was calculated by subtracting generalizability error from total error.

Table 9 shows the total, estimation and generalizability errors in the sample for all three grades and all three years that the Reading First Impact Study was fielded.[14] The proportion of variation that can be attributed to true variability among the sites in the sample ranges from 0.04 to 0.49, suggesting that in some instances error caused by generalizability is quite sizable.[15] The sample of schools selected for the Reading First Impact Study was relatively homogeneous across sites — all were low-achieving, mostly high-poverty schools. Thus, in a national probability sample, it is likely that the true variability across sites would be even higher. So researchers should use considerable caution when utilizing estimated intra-class correlations from previous studies in determining sample size requirements for new studies, especially when the estimated intra-class correlations come from populations that are quite different from the population to be included in the study.

[14] Note that the estimates of total error reported here were derived using the hierarchical linear modeling program and specifying a known Level 1 variance component, and therefore are not equivalent to what would be obtained if what was used to get an estimate of the total error were the site-by-site intra-class correlations reported in Table 8.

[15] We were not able to estimate the variance components for the third-grade 2007 reading comprehension because the models did not converge.

Further Thoughts About Uncertainty

There always will be a need for researchers to translate information about uncertainty into decisions about sample size, and this task always will require some judgment from researchers. The decisions will need to take into account the attitudes toward risk of the researcher and the research funders as well as 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 principal, 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 10th, 20th, 50th, 80th, and 90th percentiles, given whatever information is available to quantify existing uncertainty.[16] 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 80th or 90th percentile of the projected distribution. Of course, this would be possible only if the resources to do so were available.

Conclusions

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 provides new empirical information about variance parameters that influence the precision of impact estimates, presenting intra-class correlations for three-level rather than two-level models and for outcomes other than standardized test scores.

The findings suggest that

1) School-level intra-class correlations are generally larger than classroom-level intra-class correlations.

2) The intra-class correlations of traditional academic outcomes are generally larger than for academic-related outcomes (for example, dropout rates), emotional and behavioral outcomes, and health outcomes.

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

3) The choice of test metric for academic outcomes will have little effect on the estimate of the intra-class correlations for a sample.

Findings also suggest that, for academic outcomes, including pretest scores and demographic characteristics at the student level can provide considerable explanatory power at the student, classroom, and school levels. For nonacademic outcomes, including demographic characteristics at the student level provide little explanatory power at the student and classroom levels in our samples, but for some outcomes including demographic characteristics at the student level explains a substantial proportion of the variance at the school level.

The paper also assesses the magnitude and implications of the uncertainty that exists when estimating intra-class correlations while planning group randomized studies. Understanding this uncertainty is important because it has implications for study design that are often overlooked. Uncertainty derives from estimation or sampling error, and generalizability error (that is, true variation among different populations). Estimation error is influenced by a variety of factors, among them the size of the sample from which the estimate comes. The precision of the estimate is influenced most by the number of groups in the sample; a proportional increase in the number of groups reduces the standard error of the intra-class correlation by far more than does the same proportional increase in the number of individuals per group. The magnitude of the underlying intra-class correlation also affects precision. Given a constant sample size and configuration, the smaller the underlying intra-class correlation, the more precise the estimate will be. The potential for considerable estimation error in intra-class correlation estimates means that researchers should assess sample size requirements using not only the point estimate of an intra-class correlation (as is usually done in practice) but also the upper and lower bounds of those confidence intervals. Depending on the degree of uncertainty in the estimate and the researchers' tolerance for risk, it may be prudent to plan for a sample that is somewhat larger than that implied by the point estimate alone.

Gereralizability error, or true differences among the underlying populations, also introduces uncertainty into the estimates. Our findings indicate that, in some instances, error due to generalizability is quite sizable. Therefore researchers should use considerable caution when using estimated intra-class correlations from previous studies to determine sample size requirements for new studies, especially when the estimated intra-class correlations come from populations that are quite different from the population to be included in the study.

We hope that these small steps will move forward the current state of science of group randomized studies.

Cluster Randomized Trial Design
Table 1
Parameters Estimated from Three-Level Model

	Unconditional ICC		R-Squared		
Outcome	**School**	**Class**	**School**	**Class**	**Student**
Academic Outcomes					
Print awareness[a] (CLIMBERS)	0.308	0.016	0.580	0	0
Blending[a](CLIMBERS)	0.149	0.011	0.346	0	0
Elision [a](CLIMBERS)	0.000	0.068	NE	0	0
Expressive vocab[a](CLIMBERS)	0.055	0.091	1.000	0	0
Stanford 9 total math scaled score[b,c]	0.081	0.026	0.494	0.627	0.482
Stanford 9 total reading scaled score[b,c]	0.059	0.086	0.840	0.880	0.510
Academic-Related Outcomes					
Breakfast participation (adjusted for attendance)[b,c]	0.206	0.000	0.385	NE	0.320
Attendance[b,c]	0.000	0.060	NE	0.525	0.311
Days tardy as a percentage of number of school days enrolled[c]	0.077	0.000	0.253	NE	0.217
Stimulus discrimination: number of trials incorrect[c]	0.000	0.051	NE	-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	NE	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: number of animals and number of things to eat combined[c]	0.054	0.046	0.771	-0.068	0.033
Emotional and Behaviorial Outcomes					
PSC status, 0=nonPSC case, 1=PSC case[c]	0.000	0.000	-3.128	NE	0.021
Sum of answers to 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	NE	0.026	0.019
Oppositional behavior[c]	0.000	0.037	NE	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	NE	0.017	0.120
Health Outcomes					
Body Mass Index percentile[c]	0.000	0.000	NE	NE	0.004
At risk of overweight[c]	0.006	0.000	0.363	NE	0.002
Considered overweight[c]	0.000	0.035	NE	-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 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 School Breakfast Pilot

(continued)

Table 1 (continued)

Project database also include indicator variables for each school district in the study sample.

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

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

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

NE=not estimable.

ICC indicates intra-class correlation.

Cluster Randomized Trial Design

Table 2

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

	Pretest			Demographics[a]			Pretest + Demographics		
Outcome	**School**	**Class**	**Student**	**School**	**Class**	**Student**	**School**	**Class**	**Student**
Academic Outcomes									
Print awareness (CLIMBERS)	0.580	0	0	0.200	0	0	0.889	0	0
Blending (CLIMBERS)	0.346	0	0	-0.053	0	0	-0.010	0	0
Elision (CLIMBERS)	NE	0	0	NE	0	0	NE	0	0
Expressive vocab (CLIMBERS)	1.000	0	0	0.394	0	0	1.000	0	0
Stanford 9 Total Math scale score	0.454	0.421	0.474	0.585	0.418	0.069	0.494	0.627	0.482
Stanford 9 Total Reading scale 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	NE	0.289	0.217	NE	0.120	0.385	NE	0.320
attendance	NE	0.499	0.311	NE	0.149	0.024	NE	0.525	0.311
Days tardy as a percentage of number of school days enrolled	0.214	NE	0.195	0.113	NE	0.017	0.253	NE	0.217

SOURCES: Where indicated, data are from the CLIMBERS database; all other data are from the School Breakfast Pilot Project 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 School Breakfast Pilot Project database also include indicator variables for school districts in the study sample.

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

NE=not estimable.

Cluster Randomized Trial Design
Table 3
Calculated Minimum Detectable Effect Size 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**
Academic Outcomes									
Print awareness[a] (CLIMBERS)	0.516	0.567	0.254	0.512	0.229	0.486	0.218	0.469	0.210
Blending[a](CLIMBERS)	0.476	0.541	0.242	0.472	0.211	0.433	0.194	0.412	0.184
Elision[a] (CLIMBERS)	0.357	0.446	0.200	0.316	0.141	0.287	0.128	0.203	0.091
Expressive vocab[a](CLIMBERS)	0.372	0.453	0.202	0.320	0.143	0.313	0.140	0.221	0.099
Stanford 9 Total Math scale score[b,c]	0.184	0.380	0.170	0.323	0.144	0.294	0.131	0.274	0.123
Stanford 9 Total Reading scale score[b,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, data are from the CLIMBERS database; all other data are from the School Breakfast Pilot Project 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 School Breakfast Pilot Project database also include indicator variables for school districts in the study sample.

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

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

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

Cluster Randomized Trial Design

Table 4

Standard Error of the Estimated Intra-Class Correlation,
Given the Estimated Intra-Class Correlation, Group Size (N), and Number of Groups (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: Author's calculation based on hypothetical data and Equation 10.

Cluster Randomized Trial Design

Table 5

Standard Errors and 95 Percent Confidence Intervals for the Estimated Intra-Class Correlations, from Unconditional Two-Level Models

Outcomes	Intra-Class Correlation (ICC)	Standard Error of ICC	95% Confidence Interval of ICC	
			Lower Bound	Higher Bound
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 vocab (CLIMBERS)	0.106	0.028	0.055	0.165
Stanford 9 Total Math scale score	0.092	0.009	0.074	0.110
Stanford 9 Total Reading scale 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, calculations are based on data from the CLIMBERS database; all other calculations are based on data from the School Breakfast Pilot Project Year 1 follow-up database.

NOTES: Estimated values for the intra-class correlations were obtained from 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 School Breakfast Pilot Project database also include indicator variables for school districts (or states) in the study sample.

Cluster Randomized Study Design

Table 6

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

	MDES associated with 95% Confidence Interval of ICC		
Outcomes	**Lower Bound**	**Point Estimate**	**Upper Bound**
Academic Outcomes			
Print awareness[a] (CLIMBERS)	0.186	0.207	0.226
Blending[a](CLIMBERS)	0.230	0.284	0.329
Elision[a] (CLIMBERS)	0.126	0.146	0.163
Expressive vocab[a](CLIMBERS)	0.133	0.140	0.146
Stanford 9 Total Math scale score[b,c]	0.173	0.188	0.202
Stanford 9 Total Reading scale score[b,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 Year 1 follow-up database.

NOTES: Calculations assume 40 students per school and 50 schools in the sample.

Estimated values for the intra-class correlations were obtained from a two-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a two-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 School Breakfast Pilot Project Year 1 database also include indicator variables for school districts in the study sample.

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

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

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

Cluster Randomized Trial Design

Table 7

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

	Number of Schools Needed for MDES = 0.25		
Outcomes	**ICC = Lower Bound**	**ICC = Point Estimate**	**ICC = Upper Bound**
Academic Outcomes			
Print awareness[a] (CLIMBERS)	28	34	41
Blending[a](CLIMBERS)	42	64	86
Elision[a] (CLIMBERS)	13	17	21
Expressive vocab[a](CLIMBERS)	14	16	17
Stanford 9 Total Math scale score[b,c]	24	28	33
Stanford 9 Total Reading scale score[b,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 Year 1 follow-up database.

NOTES: Calculations assume 40 students per school and 50 schools in the sample.

Estimated values for the intra-class correlations were obtained from a two-level model of the outcome measure without covariates. Estimated values for R-squared were obtained from a two-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 School Breakfast Pilot Project Year 1 database also include indicator variables for school districts in the study sample.

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

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

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

ICC indicates intra-class correlation.

Cluster Randomized Trial Design

Table 8

First Grade SAT 10 Reading Comprehension Intra-Class Correlations, by Study Site

					95% Confidence Interval	
	ICC	**N**	**J**	**Standard Error of ICC**	**Lower**	**Higher**
Full sample	0.083	48.1	225	0.009	0.065	0.100
Site 1	0.041	42.0	12	0.025	-0.009	0.090
Site 2	0.041	65.0	16	0.019	0.004	0.078
Site 3	0.036	46.1	20	0.018	0.002	0.071
Site 4	0.018	104.4	12	0.011	-0.004	0.039
Site 5	0.164	42.3	6	0.090	-0.012	0.339
Site 6	0.092	41.4	29	0.027	0.038	0.145
Site 7	0.068	57.6	8	0.040	-0.010	0.145
Site 8	0.075	29.5	10	0.045	-0.013	0.163
Site 9	0.055	41.5	22	0.023	0.011	0.100
Site 10	0.074	80.0	22	0.024	0.027	0.121
Site 11	0.045	60.5	13	0.023	0.000	0.089
Site 12	0.208	54.3	16	0.063	0.085	0.331
Site 13	0.105	46.5	12	0.046	0.015	0.195
Site 14	0.156	46.0	6	0.086	-0.012	0.325
Site 15	0.140	36.8	21	0.044	0.054	0.226

SOURCES: Reading First Impact Study 2005 SAT10 Reading Comprehension data, Grade 1, 15 sites, 225 schools.

NOTES: Intra-class correlations are based on a two-level unconditional model with students nested within schools.
The standard errors of the intra-class correlations were obtained using equation 10.

Cluster Randomized Trial Design

Table 9

Reading First Impact Study Total Variation Decomposition by Grade and Year

Grade	Year	Total Error	Estimation Error	Generalizability Error	Proportion Due to True Variation
Grade 1	2005	0.0015	0.0009	0.0006	0.38
	2006	0.0007	0.0007	0.0000	0.04
	2007	0.0012	0.0009	0.0003	0.21
Grade 2	2005	0.0010	0.0006	0.0004	0.44
	2006	0.0013	0.0008	0.0005	0.39
	2007	0.0008	0.0006	0.0003	0.32
Grade 3	2005	0.0007	0.0005	0.0001	0.17
	2006	0.0014	0.0007	0.0007	0.49
	2007	NE	NE	NE	NE

SOURCES: Reading First Impact Study SAT10 Reading Comprehension data, Grades 1-3, 2005-2007, 15 sites, 225 schools.

NOTES: Variance components could not be estimated for the third-grade 2007 data because the models did not converge.
Estimates of generalizability error and the proprotion of error due to true variation were obtained from unconditional Hierarchical Linear Models with schools at Level 1 and sites at Level 2.
Total variation was calculated by dividing the proportion due to true variation by the generalizability error.
Estimation error was calculated by subtracting the generalizablity error from the total variation.
NE=not estimable.

Appendix A

Definition of the Multiplier M

The minimum detectable effect of a program impact estimator is a multiple M of its standard error (Bloom, 2005). Figure A.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,

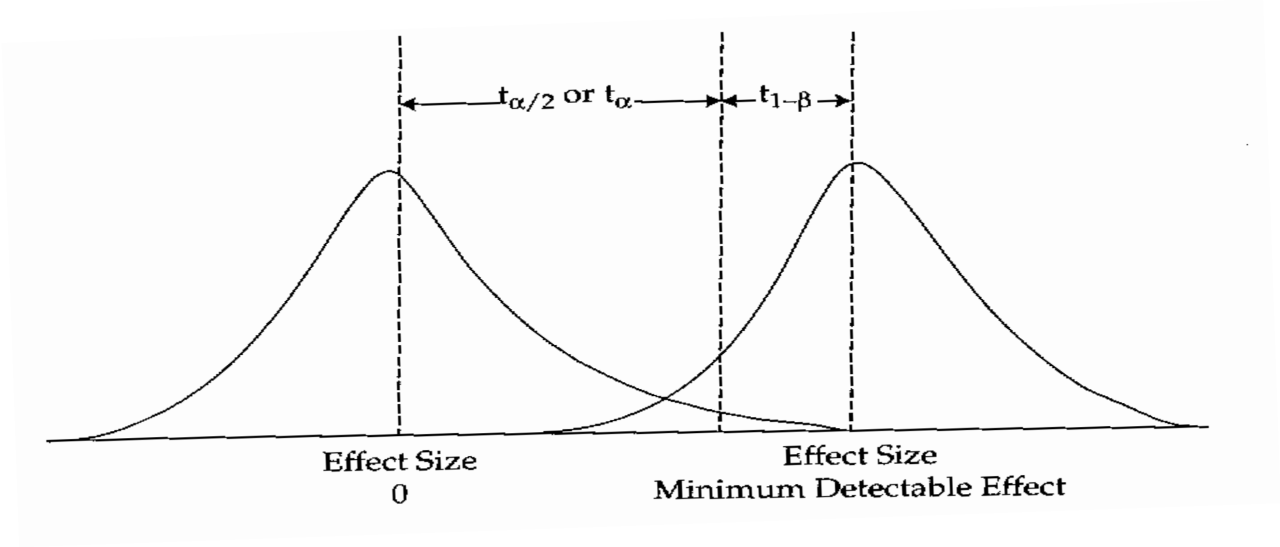
$$M = t_\alpha + t_{1-\beta} \tag{A.1}$$

For a two-tailed test,

$$M \approx t_{\alpha/2} + t_{1-\beta} \tag{A.2}$$

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 groups minus two (J-2). The multiplier for the full sample is thus referred to as M_{J-2}.

Figure A.1 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 B

Test Metrics and Estimated Intra-Class Correlation

This appendix presents estimated unconditional intra-class correlation in a variety of test score metrics for all three years and all three grades of the Reading First Impact Study data.

Cluster Randomized Trial Design

Table B.1

Estimated Unconditional Intra-Class Correlations for Reading First Impact Study SAT 10 Reading Comprehension Data Using Different Test Metrics

	Grade 1			Grade 2			Grade 3		
Metric	**2005**	**2006**	**2007**	**2005**	**2006**	**2007**	**2005**	**2006**	**2007**
Scale score	0.0827	0.0723	0.0799	0.0535	0.0676	0.0648	0.0590	0.0720	0.0656
At or above grade level	0.0630	0.0518	0.0494	0.0308	0.0487	0.0359	0.0444	0.0556	0.0449
Raw score	0.0890	0.0710	0.0796	0.0528	0.0663	0.0642	NA	NA	NA
Percentile rank	0.0841	0.0717	0.0791	0.0509	0.0673	0.0617	0.0603	0.0727	0.0649
Normal curve equivalence	0.0853	0.0727	0.0804	0.0535	0.0677	0.0651	0.0595	0.0729	0.0663
Stanine	0.0831	0.0707	0.0787	0.0530	0.0663	0.0615	0.0575	0.0716	0.0633

SOURCES: Reading First Impact Study SAT10 Reading Comprehension data, Grades 1-3, 2005-2007, 15 sites, 225 schools.

NOTES:Raw score data was not available for third grade.

Estimated values for the intra-class correlations were obtained from a two-level model of the outcome measure without covariates. All analyses include an indicator variable distinguishing treatment and control groups.

NA= not available.

References

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

American Institutes of Research, and MDRC. 2006. “Data Collection and Data Analysis Plan for a Professional Development Impact Study.” Prepared for the Institute of Education Sciences, United States Department of Education Contract No. ED-01-CO-0026/0020. July.

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 Publications.

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 *Validity and Social Experimentation: Donald Campbell’s Legacy*, Leonard Bickman (ed.), Volume 1. Thousand Oaks, CA: Sage Publications.

Carrodus, M. L., and Giles, D. E. A. 1992. “The Exact Distribution of R^2 when the Regression Disturbances are Autocorrelated.” *Economics Letters* 38: 375-380.

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.

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, Scotland: Oliver and Boyd.

Gardner, W., Murphy, J. M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., McInerny, T. K., Wasserman, R. C., Nutting, P., and Chiapetta, L. 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.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehman Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. 2008. *The Impact of two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Helland, I. S., 1987. "On the Interpretation and Use of R^2 in Regression Analysis. *Biometrics* 43: 61-69.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2 (3) pp. 172-177.

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

Moerbeek, M. 2004. "The Consequence of Ignoring a Level of Nesting in Multilevel Analysis." *Multivariate Behavioral Research* 39, 1: 129-149.

Murphy, J. M., Wehler, C., Pagano, M., Little, M., Kleinman, R., and Jellinek, M. 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. 2004. "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): 681-94.

Ohtani, K. 2000. "Bootstrapping R^2 and Adjusted R^2 in Regression Analysis." *Economic Modelling* 17: 473-483.

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

Pollitt, E., and Mathews, R. 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): 173-85.

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

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.

Reardon, Sean F. 2007. "Thirteen Ways of Looking at the Black-White Achievement Gap." Unpublished manuscript. Palo Alto, CA: Stanford University, Graduate School of Education.

Schochet, Peter A. 2005. "Statistical Power for Random Assignment Evaluations of Education Programs." Princeton, NJ: 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): 425-33.

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

Spybrook, Jessaca, H. 2007. *The Statistical Power of Group Randomized Trials Funded by the Institute of Education Sciences*. Unpublished doctoral dissertation. University of Michigan School of Education, Ann Arbor, MI.

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, pp. 1605-1614.

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

Press, S. J., Zellner, A., 1978. “Posterior Distribution for the Multiple Correlation Coefficient with Fixed Regressors.” *Journal of Econometrics* 8: 307-21.

About MDRC

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

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

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

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

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