National Title I Study of Implementation and Outcomes: Early Childhood Language Development

Part B: Collection of Information Employing Statistical Methods

May 5, 2011

MATHEMATICA
Policy Research, Inc.

Contract Number:
ED–04–CO–0112/0011

Mathematica Reference Number:
06692.602

Submitted to:
National Center for Education
Evaluation and Regional Assistance
555 New Jersey Ave.,
Capital Place, NW
Washington, DC 20208
Project Officer: Tracy Rimdzius
Contract Officer: Brenda Jefferson

Submitted by:
Mathematica Policy Research
600 Maryland Avenue, S.W.
Suite 550
Washington, DC 20024–2512
Telephone: (202) 484–9220
Facsimile: (202) 863–1763
Project Director: Christine Ross

National Title I Study of Implementation and Outcomes: Early Childhood Language Development

Part B: Collection of Information Employing Statistical Methods

May 5, 2011

MATHEMATICA
Policy Research, Inc.

CONTENTS

TABLES

PART B: COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

This submission is a request for approval of data collection activities that will be used to support the Title I Study of Implementation and Outcomes: Early Childhood Language Development (Title I ECLD). This study is being funded by the Institute of Education Sciences (IES), U.S. Department of Education (ED); it is being implemented by Mathematica Policy Research, in partnership with Decision Information Resources Inc. (DIR) and Dr. Timothy Shanahan of the University of Illinois-Chicago Center for Literacy. The study is designed to identify school programs and instructional practices associated with improved student language development, background knowledge, and comprehension outcomes from prekindergarten through third grade.

This submission is the second of a two-stage clearance request. The first submission (approved on August 2, 2010, under OMB control number 1850-0871) requested approval of the study's sampling plan, the approach to collecting the information needed to select the sample of schools, and district and school recruitment. In this package, IES is requesting approval for all data collection activities that will support the full-scale study.

Introduction and Study Overview. Reading is a critical foundational skill that enables children to learn in school and over their lifetimes. Many children, however, do not progress at the expected rate toward skilled, fluent reading that enables them to learn. The National Assessment of Educational Progress (NAEP) found that 33 percent of fourth-grade students did not

achieve a basic level of proficiency in reading in 2009 (U.S. Department of Education 2010). Children who fail to learn to read by third or fourth grade are at high risk for school dropout, with its negative implications for the trajectory of employment, income, and productivity as an adult (Crissey 2009; Rutter 1989).

Since the mid-1990s, efforts to improve reading instruction in schools and preschools serving high proportions of children at risk for reading difficulties have centered on the use of scientifically based reading instruction. Studies of these efforts show some positive effects on letter knowledge and decoding skills; fewer effects on language development in prekindergarten through first grade (Jackson et al. 2007; Judkins et al. 2008; Preschool Curriculum Evaluation Research [PCER] Consortium 2008); and no effects on reading comprehension into third grade (Gamse et al. 2008a, 2008b). Although letter knowledge and decoding are precursor skills for reading, decoding alone does not lead to comprehension (Snow et al. 1998; National Institute of Child Health and Human Development 2000; National Early Literacy Panel 2008). To increase comprehension, language development is critical, and few of the curricula and teaching strategies tested over the past decade have had a positive effect on language development.

The lack of known strategies to boost language development is important, because many children from low-income or dual-language homes arrive at preschool and kindergarten with language and literacy scores well below the average for 4- or 5-year-old children (Tarullo et al. 2008; Jackson et al. 2007; Chernoff et al. 2007). Moreover, supporting growth in young children's language development *and* background knowledge is critical if students are to comprehend text, because background knowledge is theoretically important to extracting meaning from print (Hirsch 2003, 2006;

Hoover and Gough 1990) and research evidence suggests a link between these areas of development (National Early Literacy Panel 2008).

To identify school programs and instructional practices associated with better language development, background knowledge, and comprehension outcomes for young children, ED has requested a national study. The study will focus on Title I schools because they serve substantial proportions of educationally at-risk children who enter school with language development and early literacy achievement that is below the average for children their age. It will target children from prekindergarten through third grade to measure how these outcomes may be influenced from the earliest years of formal schooling until children are first assessed in reading comprehension for school accountability purposes. To ensure that the study measures programs and instructional practices in schools with widely varying reading achievement outcomes for demographically similar children, the study will include 50 schools whose students are consistently high-performing in reading achievement outcomes and 50 schools whose students are consistently low-performing. The study will sample classrooms from prekindergarten to third grade (approximately three per grade per school). A total of about 12,000 children will be selected to participate in the study with about equal numbers per grade. Students' language development, background knowledge, and comprehension will be assessed. Principals, prekindergarten directors, and teachers will be surveyed, and parents will be interviewed. School record data will be sought for each sampled student. Analyses will estimate the associations between instructional programs and

practices and student outcomes to identify promising strategies for improving language and comprehension outcomes for educationally at-risk children in these early years of school. Future studies can rigorously evaluate the promising strategies.

The study will seek to answer the following questions about the growth of children's achievement from prekindergarten through grade 3 and its association with school programs and instructional practices:

- How do language development, background knowledge, and comprehension develop from prekindergarten through grade 3?
- What programs do the sample of schools use to support children's language development, background knowledge, and reading comprehension?
- What teacher instructional practices are associated with children's language development, background knowledge, and reading comprehension?
- What school programs are associated with greater student progress in language development, background knowledge, and comprehension?
- What instructional practices are associated with greater student progress in language development, background knowledge, and comprehension?

In addition, the study will address the following questions about the methodology for identifying high- and low-performing schools and measuring instructional practices:

- Can we accurately identify high- and low-performing schools using readily available school-level performance data and demographic information? Do schools tend to have consistently high or low performance across grades and across classrooms? Are third-grade assessment measures (typically the first year states collect standardized results) indicative of cumulative school effects in earlier grades?
- How can researchers measure instructional practices more reliably?

Study Timeline. The study began in October 2009 and is a five-year project. Activities planned for each year are as follows:

- *Year 1 (October 2009 to September 2010).* Planning and design activities, including defining and measuring consistently high- and consistently low-performing Title I schools in selected districts, identifying student assessments, developing classroom observation measures, drafting other data collection forms, and finalizing the study design.
- *Year 2 (October 2010 to September 2011).* Recruiting districts and schools, finalizing data collection instruments and training materials, and training data collection staff. We will also schedule and collect data in schools with August start dates, including sampling of classrooms and children.
- *Year 3 (October 2011 to September 2012).* Fall and spring data collection.
- *Year 4 (October 2012 to September 2013).* Analyze the data and write the report.
- *Year 5 (October 2013 to September 2014).* Revise the report and prepare restricted-use data files.

B. Collection of Information Employing Statistical Methods

B1. Respondent Universe and Sampling Methods

This study aims to estimate associations between school programs, instructional practices, and student growth in language development and comprehension outcomes from prekindergarten through grade 3 based on a sample of Title I schools with divergent student reading proficiency outcomes. Therefore, the goal of the sampling plan is to obtain a sample of 100 Title I elementary schools that include prekindergarten through grade 3, with 50 consistently high-performing schools and 50 consistently low-performing schools based on reading comprehension proficiency in grade 3. Selecting locations (sites) for the study and selecting schools within those locations were described in the first submission to OMB; these sampling

stages are summarized here. The current submission covers the next two stages of sample selection: selecting classes within schools and children within classes. Within each selected school, we plan to include as many as three classes in each of the five grades (prekindergarten through grade 3) and randomly select eight children per class. We discuss each step in the sampling process below.

Selecting Locations for the Study. Locations (a large school district or a geographically proximal cluster of two or three school districts) will be purposively selected for the study. We will include 10 diverse geographic locations for the study, a quantity that balances the goals of having a geographically diverse sample of schools and having a sufficient concentration of schools to make data collection cost-effective. Within each of the 10 locations, we will group elementary schools as high- or low-performing, identify schools that include prekindergarten through grade 3, and sample and recruit five schools from each group, for a total of 100 schools across all 10 locations. Study locations thus need to encompass large enough pools of high- and low-performing Title I elementary schools (that include prekindergarten to grade 3) for us to sample about five schools of each type. Locations will be selected to meet criteria for diversity of school performance and the number of schools meeting definitions of high- and low-performance in reading.

To find such locations, we used four criteria to identify an initial set of school districts for consideration:

1. ***A Large Low-Income Population and/or a Large Number of Title I Elementary Schools.*** Urban areas will help facilitate the cost-effectiveness of data collection for the study, because schools will be tightly clustered geographically. This criterion will also have analytic benefits, because the schools within the same district will share many policy and funding characteristics. Schools in proximal districts within a state also are likely to have many of the same policies.
2. ***A Small Number of School Districts with Which We Must Negotiate to Conduct the Study.*** Each potential location will be selected so as to minimize the number of organizations or entities from which we need to seek permission to obtain access to schools for the study, for example, by encompassing one or two school districts within a given geographic area.
3. ***Relatively High Average Reading Achievement of Low-Income Students.*** States with relatively high fourth-grade reading assessment scores on the NAEP have better average reading achievement, relative to other states.[1] Selecting districts in these states may increase the likelihood of identifying schools that perform well in developing students' early reading comprehension skills.[2] Focusing this criterion on students eligible for free or reduced-price lunch helps ensure that selected states have relatively high-performing Title I students.
4. ***A Diverse Set of Locations in the Continental United States.*** For external validity, the sites included in the study should be located in different regions of the country, reflecting different population characteristics and education policies.

School-level information on student reading proficiency and economic disadvantage was analyzed for each district to determine which districts have an adequate number of consistently high- and low-performing schools for the study. Both economic disadvantage and reading proficiency are

[1] NAEP is a nationally representative assessment of what U.S. students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history. NAEP reading assessments are conducted annually in grades 4, 8, and 12. NAEP assessments are administered uniformly using the same assessment in every location. Accordingly, NAEP results serve as a common metric for all states and selected urban districts; consistency of the assessment over time supports measurement of student academic progress over time.

[2] Most NAEP results are presented at the state level, however, results are reported for a small number of districts in the Trial Urban District Assessment. We have relied on the state-level results for the initial selection of sites, because district-level results are limited.

important factors, as schools with consistently high-achieving students might not be more effective than schools with low-achieving students if the high-achieving students are less disadvantaged. The next section discusses the definitions used to classify schools as consistently high- or low-performing. Applying these definitions to each school district (or cluster of districts) produced a set of schools meeting the definition of high-performing and a set of schools meeting the definition of low-performing, taking into account student economic disadvantage.

Based on the analyses of school performance in these districts, we set priorities among the potential sites using four criteria:

1. ***Sufficient Number of Schools Meeting High- and Low-Performance Definitions.*** A balanced study design requires five high-performing schools and five low-performing schools in each selected site. However, because refusals or changes in school performance can occur, a site with a larger number of schools meeting the definition for high and low performance is preferable to a location with fewer.
2. ***Substantial Differences in the Proficiency Levels of High- and Low-Performing Schools.*** To include schools with a wider variation in student performance, the differences in the average reading proficiency rates of schools categorized as high or low performers should be relatively large. Thus, sites with a larger difference in average proficiency rates of high-performing schools compared with low-performing schools have priority in site selection.
3. ***Similar Levels of Student Disadvantage in High- and Low-Performing Schools.*** School performance and student disadvantage are highly correlated, even within Title I schools. Therefore, to ensure that high-performing schools are not simply less disadvantaged than low-performing schools, sites with smaller differences in the average percentage of disadvantaged students in high-performing schools compared with low-performing schools receive priority in site selection.

4. ***Geographic Diversity of Sites.*** The final set of schools included in the study should be located in different regions of the country to reflect greater policy and population diversity.

School Selection. The universe of schools to be included in the study from each location encompasses public elementary schools (including magnet schools) and charter schools with prekindergarten to third grade. Charter schools may offer additional examples of innovative practices. Because individual schools must be approached for approval to conduct the study (following district approval), including charter schools will not add substantially to the recruitment burden.

To measure school-level reading achievement, we use the school-level percentage of third-grade students judged proficient on the state reading assessment. Although third-grade proficiency rates provide little information about student growth in language and comprehension outcomes across prekindergarten to grade 3, measuring reading proficiency at the end of this period summarizes the student's cumulative achievement through grade 3. Reading proficiency assessments measure the student's attainment of both the mechanics of reading (decoding and fluency) and the ability to extract meaning from text, which is built upon a foundation of language development and background knowledge. Accordingly, the student's reading proficiency level in grade 3 is a capstone measure of attainment of language development, background knowledge, and comprehension (as well as reading mechanics). Although not ideal, state reading-proficiency results at the school level are currently the best way to identify relatively high-performing schools across all states and school districts.

To identify consistently high- and low-performing schools, we used three years of grade 3 reading proficiency data, typically 2006 to 2008 (the latest data available). The definitions of school performance are based on a combination of the following criteria:

- ***Meeting a proficiency benchmark based on a three-year average***. Consistently high-performing schools are those with an average proficiency rate equal to or greater than the median of the three-year average proficient rate for Title I schools in the state. Consistently low-performing schools are those with an average proficiency rate less than the 25th percentile.
- ***Performance exceeds expectations based on student disadvantage***. This definition measures performance relative to the level of economic disadvantage, because performance and socioeconomic status are highly correlated. This performance measure is based on the difference between the school's actual proficiency rate and the expected proficiency rate for the level of disadvantage of the student population. Schools in the 80th percentile or above for the district are considered high-performing, and schools in the 20th percentile or below are considered low-performing.
- ***Meeting a threshold based on advanced proficiency (high-performing only)***. This definition focuses on the percentage of students scoring "advanced proficient" on the state's reading assessment (the highest category created by the state). Using the distribution of three-year average scores for all Title I schools in the state, schools are classified as high-performing if the school's three-year average percentage of advanced proficient students is above the state's 75th percentile for Title I schools. This criterion will be used only for districts in states that have a proficiency threshold on the state reading assessment that is below the NAEP "basic" threshold.[3]

Within districts, these performance definitions will classify the sample frame of schools for the study as consistently high-performing, consistently low-performing, or neither. We will request more recent data on school proficiency results from the districts that agree to participate in the study

[3] Bandeira de Mello et al. 2009.

(proficiency data from spring 2009 and possibly 2010), to determine whether the schools have maintained their high or low performance, or if other schools now meet the definitions. More recent years of proficiency data will be weighted somewhat more heavily, to account for the possibility that some schools may have successfully implemented strategies to improve performance. To maximize the variability in achievement among the selected schools within each location, we plan to exclude from selection those schools that do not meet the definition of consistently high-performing or consistently low-performing.

Among the high- and low-performing schools, we will identify those eligible for the study as:

- Title I schools, or schools with 40 percent or more of the students eligible for free or reduced-price lunch
- Schools that include prekindergarten through grade 3
- Schools with two or more classes in each grade from kindergarten through grade 3

Information on these eligibility criteria will be verified with school district officials. In some cases, we may need to contact charter schools directly to verify eligibility.

We will randomly select five consistently high-performing and five consistently low-performing schools within each location. We will consider implicitly stratifying by characteristics such as the percentage of English language learners and other factors important to the analysis. We plan to select backup schools within each location should schools refuse to participate or fail to meet the study criteria; however, these backups will be

selected as random replicates that are released as part of a probability sample.

Sampling Classes and Children. Within each selected school, up to three classes will be selected in each of the five grades (prekindergarten, kindergarten, first, second, and third grades). Schools that have only one class per grade (excluding prekindergarten) will be excluded from the sample. Schools that have two classes per grade will be included, but to compensate, we will sample additional classes from the same grade in other high- or low-performing schools (selecting a school in the same performance category) in the same district. This sampling strategy requires that we know the number of classes per grade in each of the ten study schools in a particular district before beginning to select the sample of classes. If a grade has more than three classes, we will randomly choose three for inclusion in the study (unless additional classes are needed, as discussed).

Broadly, eligible classes are those in which language arts is taught. These classes include traditional arrangements, whereby one teacher is responsible for all academic subjects and children are with that teacher for the entire day (excluding “special” classes such as physical education), and language arts classes, which may be used in schools with multiclass schedules or pullouts for children with special needs. For classes taught by the same teacher (for example, morning and afternoon kindergarten classes), we will sample one class per teacher, if there are more than three classes per grade.

Within each selected class, we plan to randomly select eight children in the fall of 2011, with the expectation that six of these students will be

eligible and receive parental consent, and five of them will remain in the school by spring of 2011. Children in selected classrooms will be eligible for the study regardless of primary language or disability, unless a child has an individual education plan (IEP) that specifies he or she cannot be assessed or a disability that our child assessment protocol cannot accommodate (for example, the child is blind or deaf). When we select more than one child from the same family—whether in the same class, grade, or school—we will randomly subsample one sibling for the study, so that only one child per family is included in the sample. The two extra children originally selected in each classroom are expected to account, on average, for those selected children who turn out to be ineligible, those who are randomly excluded siblings, and those for whom parental consent is not obtained.

Table 1 shows the target number of schools, classes, and children to be included in the sample. Section B2 discusses the power of this sample to estimate relationships between teacher and school practices and children's growth.

Table 1. Number of Schools, Classes, and Children in the Study

Locations Selected		10		
School Performance Type		High	Low	Total
Schools per location		5	5	10
Total schools		50	50	100
Grades per school		5	5	5
Classes per grade		3	3	3
Total classes		750	750	1,500
Children per class	Selected	8	8	8
	Consented	6	6	6
	Retained	5	5	5
Total children	Selected	6,000	6,000	12,000
	Consented	4,500	4,500	9,000
	Retained	3,750	3,750	7,500

B2. Procedures for the Collection of Information

a. Sampling and Estimation Procedures

Statistical Methodology for Stratification and Sample Selection. As described previously and in the first OMB submission, districts will not be randomly sampled for the study but purposively selected to contribute schools to the sample that have greater divergence in reading performance given similar levels of student economic disadvantage. Within districts, schools identified as consistently high-performing or consistently-low performing that meet the eligibility criteria (include prekindergarten through grade 3; include two or more classes per grade; and have 40 percent or more students eligible for free or reduced-price lunch) will be sampled using a stratified random sampling technique with strata defined by high and low performance (based on third-grade reading proficiency level and the level of disadvantage). If a grade has more than three classes, we plan to select a simple random sample of three classes. In schools with half-day prekindergarten and kindergarten classes, classes taught by the same

teacher will be sampled first, to associate each teacher with one class, and then teachers will be sampled. Within each class, we plan to select a simple random sample of eight children.

Estimation Procedure. Estimation procedures will assess the relationship between measures of instructional programs and practices and measures of student achievement growth in Title I schools. Schools will be selected from groups of either high- or low-performing schools within a district based on their state's grade 3 reading assessments to ensure wide variability in student achievement growth and instructional programs and practices. Rather than compare students and classrooms in the high-performing group with those in the low-performing group, we will analyze the correlations between instructional practices and outcomes using the full sample, because instructional practices across classrooms within a school are likely to vary, and examining this within-school variation is also of policy interest.

Practice-achievement relationships will be estimated using a value-added, hierarchical linear model (HLM) analysis that links student achievement with practices and programs in the study schools and classrooms, while properly accounting for the nested structure of the data, prior student achievement, and other confounding factors such as family literacy practices. The analytic goal will be to estimate a regression coefficient for instructional practice measures that indicates the correlation between higher levels of instructional practices and higher levels of student achievement or, in the case of multidimensional measures of instructional

practices, to estimate the proportion of reading achievement growth that can be explained by the *collection* of instructional practice measures, captured by the regression R-square (R^2).

The analysis will focus on estimating the relationship between instructional programs and practices and student achievement growth within each grade. The sample includes n schools (indexed by i), with c classrooms of a particular grade per school (indexed by j), and m students per class (indexed by k). Student achievement test scores will be obtained in the fall and spring of the school year so that the analyses can use student test score gains as the dependent variable. Using gain scores (or spring test scores with pretest scores as covariates) improves the precision of the estimated relationship between the mediator and student achievement. Continuous measures of instructional practices will be collected during the school year. The relationship between student achievement and measures of instructional practices can be modeled using the following framework:

(1) $y_{ijk} = \gamma_0 + \gamma_1 M_{ij} + (u_i + \theta_{ij} + \varepsilon_{ijk})$,

where y_{ijk} is the gain score for student k in classroom j in school i; M_{ij} is the observed mediator for teacher j in school i and is linked to each student in that classroom; γ_0 is the intercept; γ_1 is the relationship between the mediator and student gain scores; u_i are the school-level errors and are independently and identically distributed (IID) $N(0,\sigma_{uy})$; θ_{ij} are IID $N(0,\sigma_{\theta y})$ classroom-level errors; and ε_{ijk} are IID $N(0,\sigma_{\varepsilon y})$ student-level errors. The error

terms at each level, u_i, θ_{ij}, and ε_{ijk} are distributed independently of each other.

A test of the null hypothesis $H_0 : \gamma_1 = 0$ (no relationship between the mediator and the gain scores) could be formulated as an F test using the statistic:

(2) $t^2 = \hat{\gamma}_1^2 / \widehat{Var}(\hat{\gamma}_1)$,

where $\hat{\gamma}_1$ is an estimator for γ_1. The test will reject H_0 at significance level α if $t^2 > F_{1-\alpha}(1, n-1)$ at a specified level of power.

Degree of Accuracy Needed for the Purpose Described in the Justification. The sample size requirements for the study were developed by identifying the numbers of schools, teachers, and students necessary to address the study's main research questions with a reasonable degree of precision. The precision standards used by the Department of Education require that hypothesis tests be conducted with a significance level of 5 percent.

Our focus is on estimating a regression coefficient for instructional practice measures that provides a measure of the correlation between higher levels of instructional practices and higher levels of student achievement or, in the case of multidimensional measures of instructional practices, to estimate the proportion of reading achievement growth that can be explained by the *collection* of instructional practice measures, captured by the regression R-square (R^2). Thus, the study requires a sample with

sufficient precision to detect policy-relevant relationships between instructional practice measures and student achievement, should they exist.

An extensive previous literature has discussed the calculation of statistical power for a variety of hypothesis tests, including the statistical significance of the multiple correlation coefficient, or R^2 , estimated from a regression (see Cohen 1988; Kraemer and Thiemann 1987; MacCallum et al. 1996; and Rogers and Hopkins 1988). However, this literature does not address statistical power for hypothesis tests about the regression coefficient when data are clustered, as they typically are in education research studies such as this one. Moreover, these calculations do not account for the effects of measurement error for measures of instructional practices (Raudenbush et al. 2007, 2008, 2009). Schochet (2009) demonstrates that when statistical power calculations are adjusted for both clustering and the levels of measurement error estimated by Raudenbush et al. (2008), estimates of the relationships between instructional practices and student achievement are very imprecise. Schochet's paper also provides a method for calculating the precision of the sample to detect these relationships.[4] Raudenbush et al. (2007, 2008, 2009) provide estimates of specific sources of measurement error that yield implications for study design decisions that can reduce overall measurement error by addressing its largest components.

[4] Although Schochet's paper is primarily about the calculation of statistical power for testing hypotheses about linking a mediator (instructional practices, measured with error) with impacts on student outcomes in clustered randomized controlled designs in education research, it begins with the simpler case of estimating statistical power for testing hypotheses about these relationships in the control group. We use this simpler case as the basis for estimating the level of precision and power in estimating correlations between instructional practices and student achievement in the Title I ECLD study.

We discuss the framework to be used for the statistical power analysis that takes into account the clustered sample design and the problem of measurement error in the instructional practice measures, as presented in Schochet (2009); we discuss how decisions about the design of the observational measure can reduce measurement error to improve the power; and we present calculations of the expected minimum detectable size of correlations between instructional practices and student achievement under the current study design parameters. Finally, we identify the ways in which the study design will reduce measurement error and thereby improve the power of the study to detect relationships between instructional practices and student achievement.

Analytic Framework for the Power Analysis. We make several simplifying assumptions to focus on the essential points to demonstrate the precision of the sample design, using Equation (1) above as our analytic model. Ultimately, our analytic models will include multiple classroom practice measures as well as control variables for teacher background and family background; for simplicity, the model discussed here includes a single mediator and no covariates.[5] The assumption of the independence of M_{ij} with u_i, θ_{ij}, and ε_{ijk} implies three conditions: that the model contains no omitted variables correlated with the mediator; that the mediator and student gain scores are not determined simultaneously (meaning teachers with the

[5] The single mediator can be viewed as a collection of mediators, or a scale score derived from a set of mediators. We use just one for these analyses because the calculations shown here become intractable if multiple mediators are used, as each expression would need to include correlations among the mediators. The ideas presented provide the intuition for the calculations of the precision of the sample that can be generalized to models with multiple mediators.

highest levels of practices do not teach the best students); and that the mediator is measured without error. We discuss relaxing the third condition in the next subsection.

Schochet (2009) shows that, under these assumptions, the asymptotic variance of the ordinary least squares (OLS) estimator, $Var(\hat{\gamma}_1)$ is

(3) $$AsyVar(\hat{\gamma}_{1,OLS}) = \frac{\sigma^2}{nm\sigma_M^2}[1+\rho_1(m-1)+\rho_2(m\psi-1)]$$,

where ρ_1 is the classroom-level intra-class correlation (ICC); ρ_2 is the school-level ICC; and $\psi = \sigma^2_{Mb} / \sigma^2_M$, or the variance of the school-average mediator as a proportion of the total variance of the mediator.

Using the asymptotic variance in Equation (3) and two important relationships between the R^2 value from a regression of y on M ($R^2_{y,M}$) and the population variances and covariances of *y* and *M*,[6] the noncentrality parameter that is required to calculate statistical power can be expressed as follows:

(4) $$\delta_{OLS} = \frac{\gamma_1^2}{AsyVar(\hat{\gamma}_{1,OLSa})} = \frac{n\sigma_M^2 R^2_{y,M}}{(1 - R^2_{y,M})deff},$$

where $deff = 1 + \rho_1(m - 1) + \rho_2(\sigma_M\psi - 1)$ is the design effect due to clustering. The statistical power of the test can be computed using the noncentral F distribution as follows:

(5) $$Pr\left| F(1, n - 1, \delta_{OLS}) \geq F_{1-\alpha}(1, n - 1) \right|.$$

These formulas can be used to calculate the minimum detectable $R^2_{y,M}$ values for a given sample size, power, and significance level. However, before discussing these calculations, we first discuss how the reliability of the observational measures of instructional practices—or error in the

[6]The key relationships are (1) $R^2_{y,M} = \gamma_1^2 \sigma_M^2 / \sigma_y^2$ and (2) $\sigma^2 = \sigma_y^2(1 - R^2_{y,M})$.

measurement of these practices—affects the relationships described in Equation (4).

Reliability of the Observational Measures. Recent work by Raudenbush and colleagues (2007, 2008, 2009) indicates that the reliability of observational measures of instructional practices may be quite low, reducing the power to detect impacts of an education intervention. Schochet (2009) builds on this insight to discuss how measurement error in the observational measures of instructional practices also diminishes the precision of estimated relationships between classroom practices and student growth. Ideally, the observational measures for a given domain would reflect the true, underlying performance of the teacher on that domain. However, many additional factors can generate a large variance in the performance of teachers on an observational measure of instructional practices. Practices vary both within the day and between days, depending on activities, how students are behaving, and many other idiosyncratic reasons (such as sickness and weather). Raters can also vary in what they notice and how they judge the practices they observe.

Raudenbush et al. (2008) describe a theoretical model that includes all possible error variances that can arise from fluctuations in observed quality across classrooms (c), raters (r), segments (s), and days (d), with raters and segments nested within classrooms and days:

(6) $$Y_{rs(cd)} = \mu + \alpha_c + \beta_r + \gamma_d + \pi_{s(cd)} + (\alpha\beta)_{cr} + (\alpha\gamma)_{cd} + (\beta\gamma)_{rd} + (\beta\pi)_{rs(cd)} + (\alpha\beta\gamma)_{crd},$$

where $Y_{rs(cd)}$ is the score on the observational measure for classroom *c* by rater *r* for segment *s* on day *d*; μ is the average score across all observations; α_c, β_r, γ_d, and $\pi_{s(cd)}$ are random effects associated with classrooms, raters, days, and segments, respectively; and the remaining terms are interaction effects. The true quality score for the classroom is $\mu + \alpha_c$, while the remaining terms are measurement error arising from differences across raters (inter-rater reliability), practice variation across days and time segments within a day, and a rater's own variability across time segments and classrooms.

Raudenbush et al. (2008) used data from the Multi-State Study of Pre-Kindergarten (MSSPK) (Clifford et al. 2005; Pianta et al. 2005), in which observational assessments of classrooms were conducted using the Classroom Assessment Scoring System (CLASS) (Pianta et al. 2008) to estimate several of the sources of variation across days of observation, raters, classrooms, and observed segments within an observation day. Some of the sources of variation could not be estimated separately because, unless the design of the observation study includes a sufficient overlap of classroom observation times across raters, only some of the interaction terms are identified. Table 2, column 1, shows the variance components estimated by Raudenbush et al. (2008). The estimates suggest that the reliability of observation-based instructional practice measures might be quite low—that is, only a small fraction of the variance in the observation measure (11 percent) is actually due to variance in the true underlying quality, while a

large share of the variance is attributable to raters (60 percent). However, the estimates of error variance components also indicate targets of opportunity to reduce the overall measurement error of the classroom practices measures through better design of the observational measure, more explicit and intensive training of classroom observers, and the frequency and staffing of observations in the field. Thus, in Table 2 we provide

Table 2. Variance Components, Study Design, and the Reliability of an Instructional Practices Measure

	Raudenbush et al. (2008) Variance Component Estimates				
	Typical Number of Observations and Raters for a Large-Scale Study	Increase the Number of Segments Observed	Increase the Number of Days Observed	Increase the Number of Raters per Class	Increase Inter-rater Reliability
Variance Components					
Class (main effect)	0.11	0.11	0.11	0.11	0.20
Day	0.07	0.07	0.07	0.07	0.07
Class by day	0.14	0.14	0.14	0.14	0.14
Segment	0.19	0.19	0.19	0.19	0.19
Rater	0.29	0.29	0.29	0.29	0.19
Class by rater	0.12	0.12	0.12	0.12	0.08
Segment by rater	0.19	0.19	0.19	0.19	0.13
Number of Observations and Raters					0.55
Segments observed per day	3	5.5	5.5	5.5	5.5
Days each class is observed	2	2	4	4	4
Number of different raters per class	1	1	1	4	4
Measure Reliability	**16%**	**17%**	**19%**	**39%**	**60%**

Notes: Variance components are the relative proportion of the total variance of the measure associated with each source of variability. Estimates of the variance components in columns one through four are from Raudenbush et al. (2008).

estimates of the reliability of an observational measure of instructional practices, given the variance components estimated by Raudenbush et al. (2008) and alternative study design parameters (for example, number of observations per classroom) compared with those typically used for large-scale education studies. The results show that increasing the number of observation segments in a single day of observation and doubling the number of days a class is observed lead to modest changes in measure reliability, from .16 to .19. However, increasing the number of observers who conduct the observations in a single classroom increases reliability of the measure substantially, from .19 to .39.

For the Title I ECLD study, we plan to conduct a total of four half-day observations of each classroom, with two of the observations in the fall and two in the spring. The two-hour observation window will accommodate five or six observation segments per classroom visit. To improve reliability, we will schedule classroom observations so that a different observer conducts each one, for a total of four raters per classroom. Finally, the study team is developing an observational measure of instructional practices and explicit training materials with the goal of increasing inter-rater reliability relative to past studies. The reliability estimates in column five are based on alternative variance-components estimates that assume a one-third improvement in inter-rater reliability, resulting in measure reliability of .60. We expect to achieve this level of reliability by providing labels for each scoring level on the rubric and rigorously training observers. Moreover, we will aim for exact

agreement on most items rather than agreement within one point.[7] We will pilot our observer training procedures in spring 2011 to inform the degree to which inter-rater reliability will be improved for this study.

[7] The Classroom Assessment Scoring System (CLASS) (Pianta et al. 2008), used as the basis for calculations in Raudenbush et al. (2008), labels only three of seven scoring levels and is administered by observers trained to 80 percent reliability within one point.

Following Schochet (2009), we modify equation (4) to incorporate measurement error in the instructional practices measure:

(7) $$\delta_{OLS} = \frac{ncm\lambda R^2_{y,M}}{(1 - \lambda R^2_{y,M}) deff_1},$$

where $\lambda = \sigma^2_M / (\sigma^2_M + \sigma^2_\varepsilon)$, where σ^2_ε is the measurement error in the measure of instructional practices (this error includes the sources of variation described by Raudenbush and colleagues [2008])); $deff_1 = [1 + \rho_1 (m - 1) + \rho_2 (cm\psi^{Obs} - 1)]$ is the design effect; and $\psi^{Obs} = \sigma^2_{M^{Obs}_n} / \sigma^2_{M^{Obs}}$ is the variance of the *observed* school-average mediator (measured with error) as a proportion of the total variance of the observed mediator (measured with error).

The minimum detectable $R^2_{y,M}$ value (MDR) can then be calculated by first solving for $R^2_{y,M}$ in (7) and then using the following expression:

(8) $$MDR = \frac{\delta_O \, deff_1}{\lambda (ncm + \delta_O \, deff)},$$

where δ_O is the value for δ_{OLS} that achieves the pre-specified power level set in Equation (5) for a given sample size.

Estimated Minimum Detectable $R^2_{y,M}$ Values. Our estimates of minimum detectable $R^2_{y,M}$ values represent the smallest true $R^2_{y,M}$ values for which we can reliably find a statistically significant effect, that is, a value that we are fairly certain represents a true R^2 different than 0. If we include one mediator in the model and standardize the student gain scores and the mediators to have a mean of 0 and standard deviation of 1, then the square of the estimated regression coefficient on the mediator is the regression $R^2_{y,M}$ value, or the proportion of the variance in the student gain score that is accounted for by instructional practice measures. This $R^2_{y,M}$ value is also the squared correlation between y_{ijk} and M_{ij}. For example, a study with the power to detect a minimum detectable $R^2_{y,M}$ value of .03 would have an 80 percent chance of detecting a significant association between instructional practices and student achievement if the true correlation was at least .17.

To obtain target $R^2_{y,M}$ values for this study, we use results from previous studies that suggest about 15 percent of the variation in achievement score gains across students can be attributed to *all* differences between classrooms and schools (Schochet 2009 and Chiang 2009). We use ICCs of .05 for the classroom level and .10 for the school level. ICCs vary across studies, but Chiang (2009) presents a range of estimates based on the literature and new data sources and finds the classroom-level ICCs range

from .02 to .l5, while school-level ICCs range from .05 to .20; averaging across these sources yields ρ_1 = .05 and ρ_2 = .10.[8]

To calculate the minimum detectable $R^2_{y,M}$ given the planned sample size for the study, we assume a significance criterion of .05 and statistical power of .80, which implies δ_o = 7.84 in (8). We assume that ψ^{Obs} = .5. We calculate the estimated MDR under two assumptions about the reliability of the measures: λ = .39 assumes that we will observe the classrooms four times during the year using four different observers; λ = .60 assumes that we will also train observers to measure instructional practices more reliably, as discussed previously.

Table 3 estimates the MDR for our proposed design. We estimate that the current design—including a sample of 100 schools, observations from an average of three classrooms per grade on four days using four different observers, and assessments of five students in each class for the fall and spring—as well as variance components assumptions based on Raudenbush et al. (2008) for a

[8] These ICC estimates are lower than those presented in Hedges and Hedberg (2007) because the analyses use gain scores as the dependent variable, while Hedges and Hedberg used spring achievement scores as the dependent variable.

Table 3. Statistical Precision for Analyses Within a Single Grade, Under Alternative Assumptions About the Reliability of Classroom Measures

	Single Grade Level, Measure Reliability of .39	Single Grade Level, Measure Reliability of .60
Sample Sizes		
Schools	100	100
Classrooms per school	3	3
Students per classroom	5	5
Reliability of Instructional Practices Measure	**.39**	**.60**
Model Parameters		
Rho 1 (class)	.05	.05
Rho 2 (school)	.10	.10
Psi	.50	.50
Design effect from clustering	1.85	1.85
Estimated Minimum Detectable R-Square (MDR)	**.024**	**.016**
Minimum Detectable Correlation (square root of MDR)	**.156**	**.126**

Notes: Estimated MDRs were calculated using formulas developed in Schochet (2009) and reliability estimates from Table II.4 based on Raudenbush et al. (2008). Based on findings from previous Mathematica studies (estimated correlations control for student pretest scores), we assume a classroom-level ICC of .05 and a school-level ICC of .10. Analyses focus on one grade, with a sample of five students per class, three classes per school, and 100 schools. Estimated MDRs are for a 5 percent significance level with 80 percent power.

measure reliability of .39—would have an MDR of .024, which is consistent with detecting a correlation between the mediator and student gain scores of at least .156. Alternatively, we can gauge the size of the MDR by using as a heuristic the empirical finding that approximately 15 percent of the variance in student gain scores is attributable to teacher and school factors. Calculating the MDR of .024 as a proportion of the 15 percent of the variance in student gain scores attributable to teacher and school factors (.024/.15) suggests that the current study design can detect 16 percent of the variance attributable to teacher and school factors. Training observers to a higher level of inter-rater reliability, so that measure reliability increases to .60, would reduce the MDR to .016, which is consistent with detecting a

correlation between the mediator and student gain scores of .126. Using the proportion of student gain scores attributable to teacher and school effects, greater reliability of the instructional practices measures would enable us to detect 11 percent of the variation in student growth attributable to teachers and schools (.016/.15). Thus, improving the reliability of classroom practice measures would considerably improve the precision of the analysis.

Improving the Power of the Study. Given the sample sizes of schools, classrooms, and students for this study and the power analyses discussed here, several strategic design choices will be employed to improve the power of the study to detect relationships between instructional practices and student reading achievement:

- ***Develop observational measures of teacher practices with high inter-rater reliability.*** By developing an observational rubric with clearer guides to the ratings, observer training that is more intensive, and reliability requirements for certification that are more stringent than used in previous studies, we can improve inter-rater reliability, which is a significant contributor to the overall reliability of the classroom observation measure.
- ***Use multiple raters per class, multiple days of observation, and multiple "segments" observed to increase reliability.*** The study design calls for each observer to conduct a half-day observation, which is sufficient to observe five or six "segments," with two such observations in the fall and two in the spring. We expect to have more than one observer per site to accommodate the number of observations that have to be completed; therefore, to improve reliability, a different observer will assess classrooms at each time point.
- ***Use sensitive, reliable measures of student outcome to improve reliability of student growth measures.*** Student outcome measures will be individually administered, adaptive measures that use items appropriate for the student's achievement level, to provide an efficient and sensitive measure of growth across the distribution of student achievement.

- ***Sample high- and low-performing schools that have students from similar backgrounds.*** As a fraction of total variance in student achievement growth, the variation across classrooms and schools might then be greater than the 15 percent found in previous studies, and more of that variation might be due to the school or classroom practices. We do not reflect this variation in our power calculations because there is no empirical basis for estimating the benefits of our sampling approach; hence, we consider our MDR estimates conservative.

Relationships of the magnitudes that could be detected under just one of these strategies would be meaningful for identifying promising teaching practices to promote reading comprehension in the early grades of school. Successfully pursuing all four strategies could further improve the analytic power, and help to maximize the study's ability to detect relationships, if they exist, .

Unusual Problems Requiring Specialized Sampling Procedures. We do not anticipate any unusual problems that require specialized sampling procedures.

Use of Periodic (Less Frequent than Annual) Data Collection Cycles to Reduce Burden. We do not plan to collect data less frequently than once per year.

b. Data Collection Procedures

Data Collection Plan and Study Timeline. The Title I ECLD study includes eight complementary data collection efforts that will support answers to the study's research questions. Table 4 lists each of the activities; the grade levels, duration, and timing of each; the primary mode of data collection; and the number of responses we anticipate. A brief description of each of the activities follows:

- ***Student Assessments.*** The study will assess the language development, background knowledge, reading fluency, and comprehension of 9,000 students in the fall and 7,500 in the spring from prekindergarten through grade 3. A computer-assisted, one-on-one assessment will be administered to all students. To reduce assessment time for students in the study, the 40-minute reading-comprehension assessment for second- and third-grade students will be administered on a second assessment day. Likewise, we will administer the Spanish version of the language development assessment on a second assessment day, taking about 25 minutes. The length of the student assessment will vary by grade and home language. Across these situations, the average assessment will take a total of 55 to 60 minutes to administer to prekindergarten, kindergarten, and first-grade students, and a total of 80 to 85 minutes for second- and third-grade students, considering both days of assessment and the proportion of the sample potentially needing a second assessment day.[9] Assessors will be field staff who

[9] Students from Spanish-speaking homes will receive two language development assessments—English and Spanish. Depending on students' language skills, we anticipate the assessment in one language will be approximately 25 minutes and in the other language

complete a multiday training and pass certification on each of the measures to be used.

- ***Classroom Observations.*** Each of the 1,500 classrooms in the study will be observed in fall 2011 and again in spring 2012. Two trained observers will each observe each classroom for one half day in the fall and again in the spring, for a total of four half-day observations across the school year, using the observation measure designed for this

approximately 15 minutes, making the assessment time about 15 minutes longer for this group of students in the fall only. For students who are from homes that speak a language other than English or Spanish and who do not demonstrate English language proficiency as measured by the preLAS 2000, total assessment time will be approximately 20 minutes in prekindergarten through first grade and 60 minutes in second and third grade.

Table 4. Data Collection Activities and Duration for the Title I ECLD Study

Activity	Grades	Time Frame		Mode	Sample Size
		Fall 2011	Spring 2012		
Student Assessments	pre-K–3	60 to 85 min.	55 to 80 min.	Computer-based individual assessment	9,000 F 7,500 S
Classroom Observations	pre-K–3	Two 1/2 days per class, two observers	Two 1/2 days per class, two observers	Direct observation	1,500
Principal Survey	School	30 min.	--	Hard-copy	100
Prekindergarten Program Director Survey	pre-K	15 min.	--	Hard-copy	20
Teacher Survey	pre-K–3	--	25 min.	Web	1,500
Teacher-Student Report	pre-K–3	--	10 min. per student	Web	7,500
School Records Data	pre-K–3	--	30 min. per form[a]	Electronic forms	7,500
Parent Interview	pre-K–3	--	30 min.	Telephone	7,500

F = fall; S = spring.

-- Data not collected at this time.

[a]School record information will be collected at the end of the school year.

study. In each class, one observer will also conduct a rating the classroom reading materials, and the other observer will arrange the audiotaping of the teacher's language use during one of the spring observation sessions.

- ***Principal Survey.*** Hard-copy surveys will be administered to the principals of each of the 100 schools in the sample in fall 2011, with in-person or telephone followup to ensure high response rates. The survey will require about one half-hour of the principal's time.
- ***Prekindergarten Program Director Survey.*** Hard-copy surveys will be administered to the prekindergarten program director for questions particular to these programs that may be outside a principal's purview. This survey will include a subset of items currently on the principal survey that focus on the prekindergarten program. As with the principal's survey, it will be given to the prekindergarten program director associated with each of the 100 schools. In some school districts, a prekindergarten program director may be responsible for prekindergarten classrooms in more than one school. The survey will be completed in fall 2011, with in-person or telephone follow-up to ensure high response rates. We assume that the number of prekindergarten program directors is 20, reflecting some directors responding for all 10 schools in a district

and others responding for just two or three schools. The survey will require about 15 minutes.

- ***Teacher Survey.*** A 25-minute web-based survey will be completed by 1,500 teachers in spring 2012—about 300 teachers per grade.
- ***Teacher-Student Report.*** The study will use a web-based report to collect student-level data from teachers. These data will be collected in spring 2012. A total of 7,500 10-minute teacher-student reports will be completed. Each teacher will complete reports on approximately five students.
- ***School Records.*** The study will collect information from school records for all students in the study in spring 2012.
- ***Parent Interview.*** A 30-minute telephone interview will be conducted with 7,500 parents in spring 2012. Interviews will be conducted from Mathematica's Survey Operations Center, mostly during evening and weekend hours.

B3. Methods to Maximize Response Rates

The study design includes multiple instruments and several different respondents, and it is important to achieve high response rates for each component. To maximize response rates for the teacher-student report, the study will offer teachers the option of completing the form on the web or a hard copy. Based on past experience, we expect this approach to yield a response rate of 90 percent. The study will create a special form for the collection of school records. Schools can use this special form or submit the school records electronically. We expect to obtain school record data for 90 percent of the students in the study.

We anticipate a response rate of at least 85 percent for the teacher survey, the parent interview, the principal survey and the prekindergarten director survey. The combined use of web-based data collection and incentives encourages high responses on the teacher survey. We are requesting approval in this clearance to provide incentives to teachers and

parents for responding to the teacher survey, teacher-student report, and parent interview. We will conduct nonresponse follow-up using letters and emails and prompt respondents in person while we are in the schools conducting the student assessments.

Principals and prekindergarten directors will also be offered a multimode approach for completing the survey. They will be given the option of completing a hard copy survey or meeting with one of the study's team leaders for an interview. We will follow up by letter, email, and phone with any principals or prekindergarten directors who have not completed the survey in a timely manner.

Parent interviews will be conducted over the phone until 9:00 p.m. local time on weeknights and on Saturday and Sunday, to maximize the chance of successfully reaching a respondent at home. However, we will monitor the frequency with which calls are made to avoid alienating a potential respondent.

We expect a 90 percent response rate among eligible selected children for the fall student assessment and an 85 percent response rate in the spring. The relatively high fall student testing rate is likely because we will be on-site for several days administering tests to students, which will allow us ample time to conduct make-up sessions for students who are absent during the original test days. The lower expected spring rate accounts for any students who change schools or who have moved out of the area.

B4. Planned Testing of Procedures and Methods

The design of the survey instruments draws heavily on questions from instruments used successfully in previous research. Consequently, most of the survey questions and child assessments have been tested thoroughly on large samples, with low-income populations, and with prior OMB approval. A draft of the classroom observation protocol was pilot tested in spring 2010 and a second pilot to test training materials and the revised forms was conducted in March 2011. We have conducted a number of child assessments and parent interviews to ascertain the timings of each, and we will conduct additional pilot tests in spring 2011 with no more than nine respondents to determine problems respondents might have in providing the requested information. We have internally pretested the parent, principal, prekindergarten director, and teacher surveys to ascertain timing information and check the flow and wording of the items.

B5. Individuals Consulted on the Statistical Aspects of the Design

The following members of the study's expert panel were consulted on various aspects of the statistical design:

- Thomas Cook (professor of sociology, psychology, education and social policy, Northwestern University)
- Don Rock (consultant, formerly Educational Testing Service)
- Christopher Lonigan (professor, Florida State University)
- Christopher Schatschneider (associate professor, Florida State University)

The study sample and the plans for statistical analyses for this study were developed by Mathematica Policy Research, including Dr. Christine Ross, project director; Dr. Jerry West, principal investigator/survey director; Dr. Sarah Avellar, deputy project director; Dr. John Deke, senior researcher; Ms. Barbara Carlson, senior statistician; Dr. Kenneth Fortson, senior researcher; Dr. Peter Schochet, senior fellow; and Mr. John Hall, senior statistician.

REFERENCES

Bandeira de Mello, Victor, Charles Blankenship, and Don McLaughlin. "Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007." Research and Development Report, NCES 2010-c456. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, 2009.

Chernoff, Jodi Jacobson, Kristin Denton Flanagan, Cameron McPhee, and Jennifer Park. "Preschool: First Findings from the Third Follow-up of the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)." NCES 2008-025. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, 2007.

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

Clifford, R.M., Oscar Barbarin, F. Chang, Diane M. Early, Donna Bryant, Carollee Howes, Margaret Burchinal, and Robert Pianta. "What Is Pre-Kindergarten? Characteristics of Public Pre-Kindergarten Programs." *Applied Developmental Science,* vol. 9, no. 3, 2005, pp. 126–143.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Academic Press, 1988.

Gamse, Beth C., Howard S. Bloom, James J. Kemple, and Robin Tepper Jacob. *Reading First Impact Study: Interim Report.* NCEE 2008-4016. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2008a. Retrieved from http://ies.ed.gov/ncee/pubs/20084016/.

Gamse, Beth C., Robin Tepper Jacob, Megan Horst, Beth Boulay, and Fatih Unlu. *Reading First Impact Study Final Report.* NCEE 2009-4038. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2008b. Retrieved from http://ies.ed.gov/ncee/pubs/20094038.

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

Hedges, Larry, and E.C. Hedberg. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis*, vol. 29, no. 1, 2007, pp. 60–87.

Hirsch, E.D., Jr. "Reading Comprehension Requires Knowledge—of Words and the World." *American Educator*, vol. 27, no. 1, 2003, pp. 10–13, 16–22, 28–29, 48.

Hoover, Wesley A., and Philip B. Gough. "The Simple View of Reading." *Reading and Writing: An Interdisciplinary Journal*, vol. 2, 1990, pp. 127–160.

Jackson, Russell, Ann McCoy, Carol Pistorino, Anna Wilkinson, John Burghardt, Melissa Clark, Christine Ross, Peter Schochet, and Paul Swank. *National Evaluation of Early Reading First: Final Report.* U.S. Department of Education, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office, 2007.

Judkins, David, Robert St. Pierre, Babette Gutmann, Barbara Goodson, Adrienne von Glatz, Jennifer Hamilton, Ann Webber, Patricia Troppe, and Tracy Rimdzius. *A Study of Classroom Literacy Interventions and Outcomes in Even Start.* NCEE 2008-4029. Washington, DC: U.S. Department of Education, Institute for Education Sciences, 2008.

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

MacCallum, R., M. Browne, and H. Sugawara. "Power Analysis and Determination of Sample Size for Covariance Structure Modeling." *Psychological Methods*, vol. 1, no. 2, 1996, pp. 130-149.

National Early Literacy Panel. *Developing Early Literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy, 2008.

National Institute of Child Health and Human Development. *Report of the National Reading Panel. Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* Rockville, MD: NICHD, 2000. Retrieved August 1, 2009, from http://www.nichd.nih.gov/publications/nrp/smallbook.cfm.

Perez-Johnson, I., K. Fortson, C. Ross, C. Gentile, S. Amin, H. Chiang, and L. Campuzano. "Design Considerations for a Study to Validate Measures of Teacher Classroom Practices." Working paper. Princeton, NJ: Mathematica Policy Research, 2009.

Pianta, Robert C., Karen M. LaParo, and Bridget K. Hamre. *Classroom Assessment Scoring System Manual, Pre-K*. Baltimore, MD: Paul H. Brookes Publishing Co., 2008.

Preschool Curriculum Evaluation Research (PCER) Consortium. *Effects of Preschool Curriculum Programs on School Readiness.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Research, 2008.

Raudenbush, Stephen W. "Advancing Educational Policy by Advancing Research on Instruction." *American Educational Research Journal*, vol. 45, no. 1, 2009, pp. 206–230.

Raudenbush, Stephen W., Andres Martinez, Howard Bloom, Pei Zhu, and Fen Lin. "An Eight-Step Paradigm for Studying the Reliability of Group-Level Measures." Working paper, W.T. Grant Foundation, June 30, 2008.

Raudenbush, Stephen W., Andres Martinez, Howard Bloom, Pei Zhu, and Fen Lin. "The Reliability of Group-Level Measures and the Power of Group-Randomized Studies." Working paper, University of Chicago, 2007.

Schochet, P. "Do Typical RCTs of Education Interventions Have Sufficient Statistical Power for Linking Impacts on Teacher and Student Outcomes?" Washington, DC: U.S. Department of Education, Institute of Education Sciences, October 2009.

Snow, Catherine E., M. Susan Burns, and Peg Griffin. *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academies Press, 1998.

Tarullo, Louisa, Jerry West, Nikki Aikens, and Lara Hulsey. *Beginning Head Start: Children, Families, and Programs in Fall 2006.* Washington, DC: Mathematica Policy Research, 2008.

U.S. Department of Education. The Nation's Report Card: Reading 2009. NCES 2010-458. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010458.

MATHEMATICA
Policy Research, Inc.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC