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

Section B

December 15, 2009

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

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

Section B

December 15, 2009

MATHEMATICA
Policy Research, Inc.

CONTENTS

TITLE I RECRUITMENT OMB PACKAGE: SECTION B.....5

B. Collection of Information Employing Statistical Methods.....5

1. Respondent Universe and Sampling Methods.....5
2. Statistical Methods for Sample Selection and Degree of Accuracy Needed.....11
3. Methods to Maximize Response Rates.....15
4. Pilot Testing.....16
5. Individuals Consulted on the Statistical Aspects of the Design.....17

REFERENCES.....18

APPENDICES

APPENDIX A: Legislation Authorizing the Study

APPENDIX B: Letter to Accompany School Data Form

APPENDIX C: School Data Form

APPENDIX D: Confidentiality Pledge

APPENDIX E: Initial Letter to Districts and Notification Packet Materials

APPENDIX F: School Sample Notification Letter to Districts

APPENDIX G: Letter to Schools and Notification Packet Material

TABLES

1 Number of Schools, Classes, and Children in the Study..... 8

2 Statistical Precision Under the Current Design and with a Larger Student Sample 14

3 Statistical Precision of the Study Design..... 15

TITLE I RECRUITMENT OMB PACKAGE: SECTION B

B. Collection of Information Employing Statistical Methods

1. Respondent Universe and Sampling Methods

The goal of the sampling plan is to obtain a sample of 100 Title I elementary schools that include prekindergarten through grade three, with 50 consistently high-performing and 50 consistently low-performing schools as measured by student reading comprehension scores in grade three. The sample design must balance the cost effectiveness of data collection with adequate sample variation and precision to detect relationships between teacher practices and children's development during the school year.

To meet these criteria, we will identify 10 diverse geographic locations with large numbers of Title I students and variability in grade three reading comprehension achievement across the schools in the area. To reduce the number of different school districts in the study, the sample will be drawn from seven to eight of the largest districts with a substantial Title I population, which in most cases will include all or a large part of a major city. In addition, because some states (Mississippi, for example) have large Title I populations without having a very large city, we will select two to three states to represent such areas, as long as they have variation in reading comprehension achievement. Within each of the 10 locations, we will group elementary schools as high- or low-performing, identify schools that include prekindergarten through grade three, and sample and recruit 5 schools from each group, for a total of 100 schools across all 10 locations. We discuss each step in the sampling process below.

Selecting locations for the study. Locations included in the study should be geographically diverse, have Title I eligible schools with prekindergarten programs, and have variation in the performance of these schools (that is, high- and low-performing schools).

We will use school districts to define the urban location and focus on the largest urban areas and school districts. For states, we will combine sets of demographically similar and geographically proximal school districts to form potential locations within the state. Within these geographic areas, we will analyze school-level data on state reading assessments to ensure that the locations selected have sufficient variability in performance, controlling for the percentage of Title I students. We will examine data for additional locations should any of the original locations turn out not to meet the criteria or refuse to participate.

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. To construct a sample that is likely to include schools that vary in their effectiveness (performance), we will categorize schools along two dimensions: (1) the reading achievement of the students attending the schools and (2) the extent to which those students are disadvantaged. The study will draw on two sources of information to identify districts with sufficient variability in reading comprehension achievement, described below.

The first is the Common Core of Data (CCD), the Department of Education's primary database on public elementary and secondary schools in the United States. The CCD provides an official listing of schools and school districts in the nation and provides data on Title I status, enrollment, and student characteristics. The CCD will be used to identify districts with enough Title I elementary schools to support the selection of schools in the study.

The second source of data is state education data downloaded from schooldatadirect.org, a service of the Council of Chief State School Officers, funded by the Bill & Melinda Gates Foundation. The website provides the annual results of state third-grade reading tests at the school level. These results will be used to determine whether the district elementary schools have enough variation in performance to have "high" and "low" performers, controlling for the economic status

of their students. Schools will be categorized as consistently high or low performing based on their reading proficiency rates over several years relative to other schools in the same state (because definitions of proficiency vary widely by state). For example, schools could be categorized as "consistently high performing" if they are in the top third of study-eligible schools in their state in reading proficiency for each of the past three years. The exact definition of "high" and "low" performing will depend of the variation observed in schools and may vary from state to state.

Using the data from schooldatadirect.org, we can create four categories: (1) schools that are low achieving and more disadvantaged, (2) schools that are high achieving and more disadvantaged, (3) schools that are low achieving and less disadvantaged, and (4) schools that are high achieving and less disadvantaged. Locations selected for the study will ideally yield sufficient sample in each of the four categories to support the selection of approximately one-quarter of the sample from each category across and, to the extent possible, within the 10 locations.

To ensure that the sampled schools meet the study criteria (in particular, inclusion of preschool through grade three), we will obtain a list of Title I schools with preschool through grade three from each district. We discuss district recruiting and the request for school-level information in the next section.

With the information on Title I schools in the district containing preschool through grade three, and the achievement data from schooldatadirect.org, we will randomly select five consistently high-achieving and five consistently low-achieving schools within each location, stratifying the list of schools eligible for selection by a measure of the proportion of disadvantaged students they serve.

To maximize the variability in achievement across the selected schools within location, we plan to exclude from selection those schools that do not meet the definition of consistently high-achieving or consistently low-achieving. We plan to select backup schools within each location

should schools refuse to participate or turn out not to meet the study criteria; however, these backups will be selected as random replicates that are released as part of a probability sample.

Within each selected school, we plan to select up to three classes in each of five grades (prekindergarten, kindergarten, first, second, and third grades). If a grade has more than three classes, we will randomly choose three for inclusion in the study. Within each selected class, we plan to randomly select seven children in the fall of 2011, with the expectation that six of these will receive parental consent, and five of these will remain in the school by spring of 2011. The target number of schools, classes, and children to be included in the sample is shown in Table 1. Section B.2 below 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

Location Type		Urban		Non-Urban		Total		
Locations Selected		7*		3*		10		
School Achievement Type		High	Low	High	Low	High	Low	Total
Schools per location		5	5	5	5			
Total schools		35	35	15	15	50	50	100
Grades per school		5	5	5	5			
Total grades		175	175	75	75	250	250	500
Classes per grade		3	3	3	3			
Total classes		525	525	225	225	750	750	1,500
Children per class	Selected	7	7	7	7			
	Consented	6	6	6	6			
	Retained	5	5	5	5			
Total children	Selected	3,675	3,675	1,575	1,575	5,250	5,250	10,500
	Consented	3,150	3,150	1,350	1,350	4,500	4,500	9,000
	Retained	2,625	2,625	1,125	1,125	3,750	3,750	7,500

*We may select additional urban districts and fewer non-urban districts if we find insufficient variability in student reading achievement in the states with large Title I student populations.

District recruitment and sample frame development. Before contacting districts, we will send them a letter and notification packet that includes documents to identify the study sponsor and research team. The packet contains a list of frequently asked questions, a study fact sheet, a data collection schedule, a study brochure that provides additional details about the study, and brochures for Mathematica and DIR (see Appendix E). We will also explain the purpose and importance of the study; list the data collection activities, respondents, and burden estimates that participation will entail, along with incentives that will be provided in recognition of the burden imposed on the schools, teachers, students, and parents; convey appropriate guarantees regarding data confidentiality; highlight potential benefits to students, schools, districts, and education policymakers; and describe informative briefs on the overall study findings that we will provide them when the study is completed.

Followup phone calls will be made a few days after the notification packets have been sent. While we will not use a call script during the initial calls with the districts, topics presented during the call will include a brief description of the study, the benefits of participating, and the data collection plan. We will address any concerns districts have about the data collection procedures and the study in general. Whenever conditions look unfavorable, we will arrange a meeting with district personnel, such as the superintendent (or a designee), and anyone else the superintendent indicates, such as a top human resource official, the person responsible for district research and assessment, school board members, or a teachers union representative.

If the district requires approval by an institutional review board (IRB), we will prepare and submit all necessary materials to secure such approval. Once the district has agreed to participate in the study, we will ask them to sign an agreement noting the responsibilities of both Mathematica and the district and identifying contact people to ensure smooth communication during the study.

Once the agreement is signed and we have secured the approval of the district IRB, if such approval is required, the district official will be asked to supply information about schools in the district that will support sampling. This information is necessary because the CCD may be missing information in relevant areas. For example, prekindergarten programs that are not supported by local and state education funds are sometimes excluded from the data, such as the programs in California that are funded through First Five tobacco settlement money.

To ensure that our information about the grade levels and presence of preschool is accurate, we will request a list of schools in the district that include preschool through grade three and the number of students and classes at each grade level. Appendices B and C show the district letter and the School Data Form to be completed.

The list of schools from the districts will be matched with the data on reading achievement and Title I student population to create sampling frames within each district or group of districts.

School recruitment. Sampling of schools will proceed as discussed above and in B.2. Once schools are selected, we will send notification letters to the district identifying those schools and asking permission to contact them regarding their participation (see Appendix F). We will call the school district (the superintendent or research director) to discuss the selections and confirm permission to speak directly with the schools.

In early 2011, we will send letters and information packets to each of the sampled schools (see Appendix G). The letter will identify the study sponsor and research team, explain the purpose of the study, notify schools of the district's approval of the study, outline data collection activities, highlight the benefits of participation, and provide contact information for the study team. The information packet will include the same materials found in the district's notification packet (a list of frequently asked questions, a study fact sheet, a data collection schedule, a study brochure, and brochures for Mathematica and DIR).

The letter will be followed up with a phone call to the school principal. During the call we will describe the study in more detail, answer any questions the principal may have about participation, and identify the principal's designated point of contact at the school.

Once a school agrees to participate, we will discuss possible approaches for ensuring teachers are supportive of the project and the approach that will be used to secure the consent of the study students' parents. In general, we will attempt to contact principals directly, but we recognize that some districts may want to deal directly with their own schools. We will formally confirm schools' commitments to cooperate with the study by requesting principals to sign an agreement by May 1, 2011, at the latest.

2. Statistical Methods for Sample Selection and Degree of Accuracy Needed

As described above, districts will not be randomly sampled for the study. Within districts, schools that meet eligibility criteria (level of disadvantage, inclusion of preschool through grade three, and level of third-grade reading comprehension proficiency) will be sampled using a stratified random sampling technique with strata defined by third-grade reading proficiency level and the level of disadvantage. Assuming most classes within a grade within a school are about the same size, if a grade has more than three classes, we plan to select a simple random sample of three classes. Within each class, we plan to select a simple random sample of seven children.

Statistical power considerations. Based on existing research on the relationship between classroom practices and test score gains, meaningful effect sizes are likely to be modest; thus, to detect policy-relevant relationships, sample sizes must be relatively large. A broad, composite summary measure of classroom practices could conceivably explain a large proportion of the variance in test score gains if it covers a wide range of classroom factors that affect test scores. However, any individual component (or subscale of related components) will, by definition, explain a much smaller proportion of the variance. Below, we describe two important considerations in

specifying a power analysis and follow with estimates of the statistical precision of our proposed design.

Measurement error. Two recent studies identify this as an underexplored but potentially serious issue with observational measures of classroom quality (Perez-Johnson et al. 2009; Schochet 2009, draft submitted to IES). Ideally, the observational measures for a given domain would reflect the true, underlying performance of the teacher on that domain. However, many factors can generate a large variance in the performance of teachers in an actual observation measure. 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 will also vary, although they will be trained to achieve a stated reliability standard.

Estimates by Raudenbush et al. (2007) suggest that the reliability of observation-based quality scores may be quite low—that is, only a small fraction of the variance in observation measures is actually due to variance in the true underlying quality. They find that the reliability of the quality scores can be improved by increasing the number of days and time per day that a classroom is observed. Accordingly, the measurement plan calls for fall and spring classroom observations, each with two observers rating teacher practice for a half day.

- **Clustering of students.** Students will be clustered within classrooms, classrooms within schools, and schools within the selected locations. Because students within the same classroom are exposed to the same idiosyncratic conditions and, hence, are likely to have correlated outcomes, they cannot be considered statistically independent. This must be accounted for in the analysis and has the effect of reducing the effective sample size.

For these reasons, it is critical that the sample sizes are adequate to meet the needs of the study. For ease of interpretation, we calibrate our precision estimates based on R-square values—that is, the proportion of variation in student gain scores that can be explained by a particular teacher practice measure.

Our estimates represent the smallest true R-square value for which we can reliably find a statistically significant correlation. In classroom settings, our estimates suggest that 15 percent of the variation in test score gains can be attributed to *all* differences between classrooms, schools, and PSUs. However, for any given measure, the proportion is likely to be much smaller. For example, if the R-square for a given measure is 3 percent, this represents 20 percent of the total variation in test score gains that can be attributed to schools and classrooms.

Table 2 estimates the minimum detectable R-squares (MDRs) for our proposed design, based on the methodology developed by Schochet (2009). We estimate that the current design that includes observing an average of three classrooms per grade, preschool through grade three, all classrooms twice during the year, and assessing five students by spring in each class, would have an MDR of 1.7 percent for the full sample and 3.1 percent for grade-specific analyses (second column). The design would also have MDRs of 2.0 and 2.5 for subgroups comprising 50 and 25 percent, respectively, of the students—for example, subgroups defined by race/ethnicity, free and reduced-price lunch status, or limited English proficiency. Because of the clustered design, subsampling five students within each classroom does not greatly increase the MDRs relative to using all or most students in the class (see column 1). The MDR for the full sample only improves slightly, to 1.5 percent.

As noted above, the proportion of total variation in children's growth that can be attributed to schools and teachers overall is typically about 15 percent (Schochet 2009). Therefore, if the study has the power to detect a relationship (R-square) of 1.7 percent, then it can detect practices that are associated with just over 10 percent (1.7/15) of the total variation that can be attributed to teachers and schools. For single-grade analyses, where the MDR is 3.1 percent, the study can detect practices associated with just over 20 percent (3.1/15) of the total variation that can be attributed to teachers and schools.

Table 2. Statistical Precision Under the Current Design and with a Larger Student Sample

	Larger Student Sample Design	Current Design
Sample Sizes		
Schools	100	100
Classrooms per school (all grades)	15	15
Students per classroom	15	5
Total classrooms	1,500	1,500
Total students	22,500	7,500
Data Collection Parameters		
Raters per segment	1	1
Days observed for each class	2	2
Segments observed per day	6	6
Measure reliability	30%	30%
Estimated Minimum Detectable R-Square		
Full sample	1.5%	1.7%
Single grade-level analyses	2.2	3.1
Two adjacent grade levels	1.8	2.2
50% subgroup of students	1.6	2.0
25% subgroup of students	1.8	2.5

Notes: Estimated MDRs calculated using formulas developed in Schochet (2009) and reliability estimates from Raudenbush et al. (2007). Based on findings from previous Mathematica studies, we assume a classroom-level intraclass correlation (ICC) of 0.05, and a school-level ICC of 0.10.

There are a number of considerations that could improve the power of the study. First, analyzing two grade levels together when there are common student outcome measures (for example, preschool and kindergarten) improves the power considerably relative to the single-grade analysis, to a minimum detectable R-square of 2.2 percent, or 15 percent of the total variation attributable to teachers and schools. In addition, two particular features of this study may improve the power to detect relationships between classroom practices and children's growth, making our assumptions based on the research to date more conservative than what we will experience in this study:

1. The sample will include high-performing and low-performing schools but with students from similar backgrounds. Thus, as a fraction of total variance in student achievement growth, the variation across classrooms and schools may be greater than the 15 percent found in previous studies and more of that variation may be due to the school or classroom practices. We do not reflect this in our power calculations because there is no empirical basis for estimating the benefits of our sampling approach.

2. Observational measures will be refined to better address the goals of this study: to identify classroom practices associated with children's language development, background knowledge, and comprehension outcomes. This will mean dropping areas that are of less interest to this study to save observer time to go into areas of interest (language development, comprehension) in a more detailed way.

In summary, the study design with 100 schools, 1,500 classrooms, and 9,000 (fall 2011) or 7,500 (spring 2012) children is powered to detect relationships (R-square) of 3.1 percent in single-grade analyses, or practices associated with just over 20 percent (3.1/15) of total variation attributed to teachers and schools within one grade. Combining two grades at a time would improve power so that relationships of 2.2 percent can be detected, or practices associated with 15 percent (2.2/15) of total variation attributed to teachers and schools within two grades. Relationships of these magnitudes would be meaningful for identifying promising teaching practices to promote reading comprehension in the early grades of school.

Table 3. Statistical Precision of the Study Design

Data Collection Parameters	
Raters per segment	1
Days observed for each class	2
Segments observed per day	6
Measure reliability	30%
Estimated Minimum Detectable R-Square	
Full sample	1.7%
Single grade-level analyses	3.1
Two adjacent grade levels	2.2
50% subgroup of students	2.0
25% subgroup of students	2.5

3. Methods to Maximize Response Rates

To maximize response rates on the teacher student report, the study will offer teachers the option of completing the form on the web or on hard copy. Based on past experience, we expect a response rate of 90 percent using this approach. The study will create a special form for the collection of school records. Schools can use this special form or opt to 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 on the teacher survey, the parent interview, and the principal survey. The combined use of web-based data collection and incentives encourages high responses on the teacher survey. Teachers will also be given the opportunity to complete the survey on hard copy. We will conduct non-response follow-up using letters and e-mails, and also prompt them in person while we are in the schools conducting the student assessments.

Principals will also be offered a multi-mode approach for completing the survey. They will be given the option of a hard copy survey or meeting with one of the study's team leaders to conduct an interview. We will follow up by letter, e-mail, and phone with any principals 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, in order 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 will also offer the option of completing a hard copy of the parent interview, which will be distributed at school for students to take home. We will provide parents with postage-paid envelopes to return their completed interviews.

We expect a 90 percent response rate for the fall student assessment and an 85 percent response rate in the spring. The relatively high fall student testing rate is expected because we will be in a given site for several days administering tests to students and this 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.

4. Pilot Testing

We will pilot test measures that are new, that are adaptations and extensions of existing ones, that have limited information on reliability and validity for the population in this study, and for

which we have concerns about how measures perform combined with others. Each of the pilot tests will be conducted with no more than nine respondents. Data collection forms and the results of the pilot testing will be included in the second OMB submission in December 2010.

5. Individuals Consulted on the Statistical Aspects of the Design

The study's expert panel will be consulted when finalizing the statistical aspects of the study design. The panel will be finalized by January 2010 and a list of panel members and their contact information will be included in the second OMB submission in December 2010. 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, survey director; Dr. Sarah Avellar, deputy project director; Dr. John Deke, senior researcher; Ms. Barbara Carlson, senior statistician; and Mr. John Hall, senior statistician.

REFERENCES

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.

Raudenbush, SW, Martinez, A, Bloom, H, Zhu, P, & Lin, F. (2007). "The Reliability of Group-Level Measures and the Power of Group-Randomized Studies." University of Chicago. Working paper.

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.

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