REL
SOUTHEAST

OMB PACKAGE

2.1.1: The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (Vocab)

Supporting Statement Part B

Date Resubmitted:
July 20, 2007

Contract Number:
ED-06-CO-0028

Submitted to:
Gil Garcia
Institute of Education Sciences
U.S. Department of Education

Submitted by:
Ludwig D. van Broekhuizen
REL-Southeast
SERVE Center
915 Northridge Street, Second Floor
Greensboro, NC 27403-2112
(800) 755-3277
(336) 315-7400

TABLE OF CONTENTS
OMB Package Supporting Statement B

Appendix List

Supporting Statement for Request for OMB Approval of Data Collection/Needs Assessment for the REL-SE

Part B. Collections of Information Employing Statistical Methods

Background

This document presents Section B of the Supporting Statement for *An Evaluation of the Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten*. The intervention that will be evaluated, *PAVEd for Success* (PAVE), is an early literacy program designed to enhance vocabulary development among kindergarteners in high-poverty schools. Vocabulary skills are critical for learning to read, as they provide an essential foundation for decoding, fluency, and reading comprehension. Children who live in poverty are more likely to enter school with more poorly developed language skills, including vocabulary (Smith, Brooks-Gunn, & Klebanov, 1997), and continue to fall further behind through elementary school. Correlational research consistently finds that early oral language, vocabulary, and other preliteracy skills are related to later literacy skills, including reading comprehension (Storch & Whitehurst, 2002; Tabors, Snow, & Dickinson, 2001). The goal of intervening early to improve children's vocabulary skills is to put children who are at risk of poor reading outcomes on a trajectory toward better reading outcomes.

This evaluation is being conducted by the Regional Education Laboratory – Southeast Region (REL-SE), located at the SERVE Center, University of North Carolina at Greensboro (SERVE), and its subcontractors: Abt Associates Inc., the University of Georgia (UGA), and Empirical Education Inc. (EEI). The Regional Educational Laboratory (REL) Program is authorized under the Education Sciences Reform Act of 2002, Part D, Section 174, (20 U.S.C. 9564) and administered by the Institute of Education Sciences' National Center for Education Evaluation and Regional Assistance. (Part D, Section 174 of the Education Sciences Reform Act of 2002 is attached in Appendix A.) The priority for the REL program is to provide policymakers and educators with expert advice, training, and technical assistance, based on the latest findings from scientifically valid research, related to meeting the requirements of the No Child Left Behind Act (Institute of Education Sciences, 2007; http://ies.ed.gov/ncee/edlabs/about/). In instances where there is insufficient scientific evidence for the effectiveness of strategies to improve learning, the RELs are charged with conducting rigorous studies of such strategies. Each of the Regional Education Laboratories is directed to conduct rigorous studies designed to address issues of high priority to the region. The studies must meet IES' standards for field tests based on experimental designs and are intended to establish causally valid evidence of the effects of proposed policies, programs or practices on academic achievement or other related needs of the region.

Through extensive discussions conducted by REL-SE to determine the most pressing educational needs of the region, southeastern state department reading directors and the Director of the Florida Center for Reading Research voiced widespread agreement regarding the need for a vocabulary intervention among kindergarten students in the southeast region of the United States. The highest priority was placed on a vocabulary intervention for two reasons: (1) children in the region are well behind national averages in vocabulary skills and (2) vocabulary knowledge is an essential component of literacy development that has generally been more difficult to impact than other emergent literacy skills, such as letter knowledge.

Several psychometric studies including southeastern children suggest that poor and/or African-American children from this region may have particularly low vocabulary scores, averaging about one standard deviation below the national average (Campbell, Bell, & Keith, 2001; Restrepo, Schwanenflugel, Blake, Neuharth-Pritchett, Cramer, & Ruston, 2006). Difficulties that southeastern children have with vocabulary manifest themselves as they transition from learning to read to reading to learn. Averaging over state report cards of Alabama, Florida, Georgia, Mississippi, and South Carolina, 18% of third and fourth grade children do not meet state standards in reading. By middle school, this rate increases dramatically to 32%. The trend is far worse for African-American and economically disadvantaged children in the region, of whom 41% and 40%, respectively, did not meet state standards for reading in middle school.

A focus on vocabulary is a good place to start in providing regional access to higher reading achievement. Despite being a critical element in reading success, vocabulary is not a well-established part of kindergarten curricula or standards. In contrast, alphabet knowledge, phonological awareness, and print uses are typically part of kindergarten instruction and standards. According to teacher estimates, kindergarten classrooms in the U.S. involve approximately equal amounts of time spent on teacher-directed instruction in reading, numbers, and the alphabet (Heaviside & Farris, 1993; Guarino, et al, 2006); however, there is not much evidence that kindergarten teachers explicitly focus on vocabulary per se. Furthermore, standards that kindergarten children must meet are relatively consistent across the U.S. (Graue, 1999), typically focusing on the alphabet, phonological, and print knowledge, with vocabulary often not explicitly included.

There is very little research on programs that focus on vocabulary with kindergarten children. A few experimental studies have found short-term benefits of kindergarten vocabulary programs (Coyne, Simmons, Kame'enui, & Stoolmiller, 2004; Robbins & Ehri, 1994); however, currently there is insufficient evidence for effective strategies for this age range. The PAVE program is one vocabulary program that has shown promise, but more rigorous testing is required to establish strong evidence of its effectiveness.

Overview

This study examines the effectiveness of *PAVEd for Success* (PAVE), an intervention to promote vocabulary development among kindergarteners in high poverty schools. The overall goal of the intervention is to enhance children's vocabulary knowledge as a foundation for literacy development; however, we hypothesize that the PAVE intervention affects children's development through impacts on teaching practice. The research questions for this project focus on two major areas: (a) the impact on student's vocabulary outcomes and (b) the impact on kindergarten teachers' vocabulary and broader literacy instructional practices.

The PAVE intervention provides teachers with professional development through which they learn research-based strategies for enhancing children's vocabulary development during interactive book reading; cognitively challenging conversations; and direct vocabulary instruction. Teachers are trained to increase the number and quality of conversations with students, to engage in more active and more frequent small-group book reading, and to use explicit strategies for directly teaching vocabulary. Higher-quality teacher-child conversations involve, for example, a broader diversity of words, more rare words, and more cognitively challenging talk, than tend to be used in commonly occurring conversations about routine and concrete matters. Teachers are trained to engage in frequent and interactive storybook reading and re-reading with children, including asking cognitively-challenging questions, requesting children to predict events and draw conclusions, and making connections to children's experiences. The training provides teachers with specific skills and techniques for focusing on vocabulary in conversations and book reading in order to enhance children's learning.

The intervention and study will be conducted in a rural area of Mississippi known as the Delta, which is characterized by high poverty and low student achievement.

1. ***Describe (including a numerical estimate) the potential respondent universe and any sampling or respondent selection methods to be used. Data on the number of entities (e.g., establishments, State and local government units, household, or persons) in the universe covered by the collection and in the corresponding sample are to be provided in tabular form for the universe as a whole and for each of the strata in the proposed sample. Indicate expected response rates for the collection as a whole. If the collection has been conducted previously, include the actual response rate achieved during the last collection.***

The Mississippi Delta region was selected to be the target of this intervention because of the lagging vocabulary skills of children in the region, the extremely high poverty rate and low student achievement in the region, and a state legislative focus on revitalization of the Delta region (Mississippi House Bill 1034), which emphasizes meeting the early educational needs of children through grade 3. We will select a purposive sample of elementary schools with kindergarten classrooms in the Mississippi Delta region. The primary goal in testing the effectiveness of PAVE is to achieve internally valid comparisons by randomly assigning schools to treatment conditions. Although we would prefer and will attempt to sample ALL of the elementary schools with kindergarten in the Mississippi Delta region, we will only be able to sample schools that are willing and able to participate, and that provide good matches. We cannot force schools to participate. Although the sample does not have statistical generalizability to the schools in the Mississippi Delta, the study design and sample enable us to achieve internally valid comparisons through random assignment of schools to treatment conditions. In addition, through data provided to us by the Mississippi Department of Education, we will be able to make some limited inferences about the representativeness of our sample to schools in the Mississippi Delta region and to all schools in Mississippi.

The universe of districts, elementary schools, kindergarten teachers, and kindergarten students in the Mississippi Delta region is shown in Table B.1. In the Delta region, there are a total of 33 school districts and 84 schools that include kindergarten[1]. The 33 school districts and 84 schools

[1] These 84 schools all include at least *one* kindergarten classroom; however, we will be restricting our sample to schools with at least *two* kindergarten classrooms. At this time, we do not know how many schools have only one kindergarten classroom. A rough estimate based on the school's total enrollment and total number of grade levels suggests that there may be approximately four schools with only one kindergarten, but we still have to determine for certain the number of schools with at least two kindergarten classes.

constitute all the school districts and all the elementary schools with kindergarten in the 16 counties of the Delta region. Based on 2006-2007 enrollment, there are 286 kindergarten teachers and 5,729 kindergarten students in the respondent universe.

Table B.1
Universe and Proposed Sample

	Universe	Sample Target
Districts	33	33
Schools	84	84
Teachers	286	168
Students	5,729	1680

There are three criteria for inclusion in the study, which may limit the universe slightly, but we have not yet determined precisely how many schools meet the criteria. To be included, schools must:

- Be located in a high-poverty community;
- Have at least two kindergarten classrooms; and
- Serve predominantly low-income children (based on eligibility for free and reduced-price meals);

To participate in the study, schools must also:

- Be willing and able to allow two consenting teachers to be selected for the study at random; and
- Be willing and able to have 10 randomly selected students from each selected classroom participate in the study.

In addition, both districts and schools must:

- Be willing and able to cooperate with the data collection and logistical needs of the evaluation; and
- Agree to the random assignment of schools to treatment and control conditions.

We anticipate a high level of state and district support for the study, which will help us achieve a high participation rate among schools. Furthermore, we anticipate that schools will be motivated to participate because they want to receive the intervention. Vocabulary improvement is an explicit educational standard in Mississippi; however, few specific instructional strategies for achieving that standard are currently in place. Despite strong state and district support combined with eagerness for the intervention among schools, we cannot expect every school in the universe to be willing and able to participate in the study. We estimate that the participation rate for schools will be approximately 80%.

If, from the universe of 84 schools, 80% participate, we will have a sample of 67 schools from the Mississippi Delta region. If the participation rate is lower, we will add neighboring and demographically similar districts to the universe one at a time until we obtain a sample of at least

60 schools. Each time we add a district to the universe, all elementary schools with kindergarten in that district will become part of the universe, and we will attempt to include all of them in the sample. Once we have sampled at least 60 schools, we will no longer add new districts to the universe, but we will continue to sample all willing and able schools from districts already included in the universe. Based on a power analysis (described in Section B2d), we have determined that we will require at least 60 to 80 schools in the sample to detect impacts on students of the size anticipated based on previous research on PAVE. Although we have identified this target range for attaining sufficient power, there is no target sample size after which we will stop attempting to recruit schools in the universe. In other words, even if we have recruited 80 schools, we will continue to try to recruit the additional schools in the universe that have not declined to participate. We will include in the sample every willing and able school in the universe.

We will obtain data on school characteristics (e.g., number of students, racial/ethnic composition; percent of students receiving free- and reduced-price lunch; state proficiency test scores) from the Mississippi Assessment and Accountability Reporting System (MAARS), a searchable online database on the Mississippi Department of Education website http://orsap.mde.k12.ms.us:8080/MAARS/indexProcessor.jsp). Information on school characteristics will be used for determining eligibility and for blocking schools prior to random assignment.

From the pool of eligible schools that are recruited and have signed consent forms, we will block schools into groups based on their previous experience with reading initiatives (e.g., Reading First). Because we expect experience with other vocabulary or reading initiatives to be associated with differences in teachers' baseline instructional practices, we want to ensure that the treatment and control groups are equivalent in this regard. To obtain information about vocabulary/reading initiatives that are currently being implemented at the district or school level, we will talk with personnel from each school district. We do not yet have an estimate of the number of schools per strata. Within reading initiative blocks, we will further block schools on observable characteristics (e.g., percent of students receiving free- and reduced-price lunch; state proficiency test scores) and then randomly assign schools to either the intervention or control condition.

Although it is common to block schools by district prior to random assignment, we will be forming cross-district blocks. Were we to block by district, we would only be able to include districts with two or more elementary schools with kindergarten. However, 16 of the 33 districts in the Delta region (48%) have only 1 elementary school with kindergarten. Consequently, blocking by district would result in a substantial loss of schools that could be included in the sample. Furthermore, even within districts with 2 or more schools, schools may be too dissimilar for randomization to produce balanced samples in each condition. With few schools in a district, randomization could result in non-comparable treatment and control groups. For these reasons, we decided to block schools by a more substantively relevant variable than district (i.e., prior experience with a reading initiative). We expect that schools' experience with other vocabulary or reading initiatives is more strongly associated with differences in teachers' baseline instructional practices than school district is. In addition, blocking by experience with reading initiatives will enable us to have larger strata, thereby helping to ensure that randomization will result in more comparable treatment and control groups.

For both teachers and students, we will select a random sample after obtaining consent or parental permission, respectively. We will ask every kindergarten teacher in the school to consent to participate. From the pool of consenting teachers in each school, we will randomly select two teachers. Regardless of whether we were to select teachers prior to or after their consent, we would have the same consent rate. Sampling teachers after obtaining consent is a better strategy for obtaining the desired sample size. The probability of selection will vary for each school, depending on the number of kindergarten teachers in the school; we will randomly select 2 teachers per school regardless of the total number of teachers in the school.

The sampling plan for students is similar to the plan for teachers. We will attempt to obtain parental permission for every kindergarten student in each participating classroom. From the pool of students with parental permission in each participating classroom, we will randomly select 10 students with parental permission. The probability of selection will vary for each classroom depending on the total number of students in the classroom.

Among teachers, we expect the consent rate to be 80%. Once teachers have agreed to participate and are selected for the study, we anticipate a negligible amount of attrition during the intervention year (less than 5%). Although we anticipate 20% attrition among teachers during the subsequent year, data collected from teachers in the year following the intervention will not be used to examine intervention impacts but to investigate the sustainability of intervention practices among treatment teachers only. Among students, we anticipate a response rate of 85% - 90% at the end of first grade.

2. *Describe the procedures for the collection of information including:*

2a. Procedures and Protocols for Information Collection

Student assessments. Trained data collectors, hired and overseen by staff from the SERVE Center at UNC-Greensboro, will conduct child assessments at the beginning of the kindergarten year (Fall 2008) as a pretest, at the end of the kindergarten year (Spring 2009) as a posttest, and at the end of the first grade year (Spring 2010) as a follow up. Data collectors will be independent of the intervention implementation and will be blind to treatment status. Table B.2 shows the timeline for assessing students and the data collection instruments to be used at each time point.

At each time point, data collectors will assess students' receptive and expressive language skills using the **Peabody Picture Vocabulary Test-4** (Dunn & Dunn, 2007) and the **Expressive Vocabulary Test-2** (Williams, 2007), two nationally-normed, standardized and commonly used assessment instruments. Children's responses on these assessments (and the Woodcock Reading Mastery Test-Revised/Normative Update, below) will be recorded directly on the assessment protocol. Children's raw and standardized scores will be entered into an electronic database in preparation for statistical analysis. To ensure accuracy, all data will be entered again into a duplicate electronic database. Data in both databases will be compared; all discrepancies will be identified and rectified. Each child will be assigned a unique ID number; children's names will not be entered into the database. Sample items from the Peabody Picture Vocabulary Test and the Expressive Vocabulary Test are presented in Appendices B and C, respectively.

Table B.2
Schedule of Data Collection from Students

	SY 2008-2009		SY 2009-2010
	Fall 2008	Spring 2009	Spring 2010
Student Assessments	**Pre** (Weeks 4-7)	**Post** (Weeks 23-25)	**Follow-up**
Peabody Picture Vocabulary Test-4	X	X	X
Expressive Vocabulary Test-2	X	X	X
Lexical Diversity		X	X
Woodcock Reading Mastery Test-R/NU[2]			X

At the end of kindergarten (Spring 2009) and the end of first grade (Spring 2010), the data collectors will also obtain an elicited language sample to examine **students' lexical diversity,** i.e., the number of unique words relative to the total number of words in their speech production. Children will work one-on-one with a trained staff member from SERVE and be asked to tell a "story" from a wordless picture book for about 10 minutes. Each child's narrative will be audiotaped, transcribed, and analyzed using the Computerized Language Analysis program (CLAN; MacWhinney, 2000), a widely used program for coding and analyzing language samples. Because of the resource intensive nature of this type of data collection, samples will be collected from four students per participating classroom, and the four students will be selected randomly from the total pool of participating students in each classroom. The administration protocol for the Lexical Diversity measure is attached in Appendix D.

When children are in first grade (Spring 2010), data collectors will also administer the **Woodcock Johnson Reading Mastery Test-Revised/Normative Update** (WRMT-R/NU, Woodcock, 1998) to assess reading achievement. Subtests include: Word Identification, in which children read words in a list format; Word Attack, in which children read nonsense words; Word Comprehension, with antonym, synonym, and analogy sections; and Passage Comprehension, in which children read text passages to themselves and fill in the blank to demonstrate understanding.

Training on student assessments. Data collection training will involve training on how to administer the Peabody Picture Vocabulary Test-4 (Dunn & Dunn, 2007) and the Expressive Vocabulary Test-2 (Williams, 2007) and on the protocol for obtaining an oral language sample. For the second year of data collection, assessors will also be trained on how to administer the Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU; Woodcock, 1998). Training will be held in the summer of 2008, prior to the start of the intervention year, in a location most convenient and cost-efficient for trainers and trainees. A refresher training

[2] The Woodcock Reading Mastery Test-R/NY is a copyright protected instrument, thus it cannot be included as an appendix to this document.

session will be offered prior to data collection at the end of first grade (School Year 2009-2010). Training will be led by senior Abt Associates staff with experience with these specific measures or similar early literacy measures and with training data collectors. A training manual will be prepared prior to the data collection. The training will be a combination of reviewing the instruments in detail and hands-on practice. Data collectors will also need to establish 85% reliability with the trainers on the PPVT-4, EVT-2, and WRMT-R/NU.

Trainers for both the student assessments and the classroom observations (see below) will discuss working with schools and teachers to set up times to do the child assessments, classroom observations and other data collection. Discussions will include working within the schools' space constraints; attaining child assent (for data collectors trained on child assessments); handling unforeseen issues or circumstances that might arise during data collection; building rapport with children; conducting quality checks on the data collected; protocols for transmitting data; and who to contact for help during the data collection process. Staff from Abt Associates will also discuss the procedures for Privacy Act compliance; the responsibility of each data collector to ensure confidentiality of the respondents; and how to carry out the commitment to confidentiality in handling, checking, and transmitting the data. Finally, data collectors will be trained to follow established protocols in order to work efficiently and effectively with children so that needed information is obtained within the least amount of time.

Abt Associates staff will develop a set of criteria prior to training that will be used to determine whether or not individuals trained as data collectors have met the standard required for data collection. Staff anticipates that some of the data collectors will need remedial training. In addition, staff will train more data collectors than are required, in the event that some of those trained will not meet the criteria even after remedial training. Staff from SERVE will provide close oversight of the data collectors over the course of the first weeks of data collection, to identify any problems before extensive data are collected.

Procedures for collection of student assessments. Students with signed written parental permission forms in both treatment and control classrooms will be assessed in a quiet location at the school. At each assessment point, data collectors will collaborate with children's classroom teachers to (1) identify the least intrusive times for child assessments, in order to minimize loss of instructional time and (2) employ strategies for maximizing the child's familiarity with the data collector and comfort during the assessment. The data collector will obtain the child's assent prior to beginning the collection of information. The tester will use the following script:

> *Would it be okay for you to come with me to the (*name the specific place designated by the school, e.g., library)*? I'd like to learn more about what you know about what words mean and how you tell stories. You don't have to come with me if you don't want to. You can let me know anytime when you want to go back to your room.*

If the child demonstrates fatigue during the assessments, data collectors will return the child to the classroom and complete the assessment on the following day. A protocol for administering child assessments is attached.

Classroom and teacher data. Trained data collectors, hired and overseen by staff from the SERVE Center at UNC-Greensboro, will collect classroom and teacher data in both treatment and control classrooms. Demographic information on teachers and paraprofessionals will be collected using a questionnaire (See Appendices E and F, respectively). Teachers in the treatment group will be given the questionnaire when they receive the PAVE training in the summer of 2008, just prior to the intervention school year. Intervention designers from the University of Georgia will have teachers complete the questionnaire on the day of the training. SERVE data collectors will give the questionnaire to teachers in the control group when they visit the classroom for baseline classroom data collection (see below). Completing the questionnaire with take teachers approximately 10 minutes, so they will be able to complete and return the questionnaires to data collectors during the visit.

Baseline classroom data collection will take place at the beginning of the intervention year (Fall 2008), and posttest data will be collected at the end of the school year (Spring 2009). Data collectors will be independent of the intervention implementation and blind to treatment status.

The **classroom observation** measure (attached in Appendix G) focuses on teachers' instructional practices, particularly with regard to interactions that facilitate students' vocabulary development. The instrument is a time sampling measure with 3 components: (1) every 5 minutes, observers will document the literacy content covered by teachers; (2) during the intervening time, observers will document the types and frequency of teacher talk and vocabulary instruction that occur; and (3) when teachers read books aloud to students, observers will document teachers' reading strategies, types of talk, and vocabulary instruction. Two observers will visit 15% of participating classrooms at the same time to conduct classroom observations and evaluate inter-rater reliability. Observers will attempt to be as unobtrusive and undisruptive as possible to the usual classroom occurrences. Observers will document teachers' instructional practices on a written document. Codes from the written document will be entered into an electronic database. To ensure accuracy, all data will be entered again into a duplicate electronic database. Data in both databases will be compared, and all discrepancies will be identified and rectified. Teachers will be assigned a unique ID number; their names will not be included in the database. Student ID numbers will be linked with the corresponding teacher ID, so that the electronic student and teacher databases can be linked for analysis.

In addition, every participating teacher will be audio-recorded during a 20-minute, small group instructional period focusing on literacy to examine the **lexical diversity of teachers' language** directed to students in the classroom. These samples will be collected two times during the intervention year from each participating teacher in both the treatment and control conditions, at the same time that classroom observations are conducted. The protocol for recording the literacy lesson is attached in Appendix H. The teacher language samples will be transcribed and analyzed using a language analysis computer program. Project staff will rate the overall quality of teachers' talk on recorded samples using the Teacher Interaction and Language Rating Scale (Girolametto & Weitzman, 2002). The Teacher Interaction and Language Rating Scale is attached in Appendix I.

Additional classroom and teacher data will be collected in the fall of the subsequent school year (Fall 2009) – *from teachers in the treatment group only*. Again, data collectors will be independent of the intervention, but they will not be blind to treatment status at this follow-up assessment of teachers in the treatment group.

A **fidelity assessment** will be conducted in Fall 2009 to determine if teachers in the treatment group sustain the implementation of the PAVE intervention as it was intended. For this assessment, data collectors will conduct a two-hour observation within the classroom and rate teacher's level of fidelity on key components of the PAVE program. The fidelity observation will occur during the second month of the school year in treatment classrooms only. Data collectors will be responsible for working with the schools and teachers to identify the times that, in general, are appropriate for visits and to follow established procedures. Following the fidelity assessment, data collectors will **interview teachers** about components of the PAVE intervention that they find difficult or challenging to implement. This information will be used to further refine and improve the design of the intervention. The fidelity assessment tool and the implementation challenges interview are attached in Appendices J and K, respectively.

Table B.3 shows the timeline for collecting classroom and teacher data and indicates the data collection instruments to be used at each time point.

Table B.3
Schedule for Collection of Classroom and Teacher Data

	SY 2008-2009		**SY 2009-2010**
	Fall 2008	Spring 2009	Fall 2009
Teacher Measures	**Pre** (Weeks 5-8)	**Post** (Weeks 21-25)	**Follow-up** (Weeks 5-8)
Classroom Observation[a]	X	X	
Audiotape Recorded Literacy Lesson	X	X	
Teacher Demographics Questionnaire	X		
PAVE Fidelity Assessment[b]			X
Implementation Challenges Interview[b]			X

[a] The classroom observation instrument is an observational measure designed for this study.

[b] Measures to be administered in treatment classrooms only.

<u>*Training on classroom measures.*</u> Data collection training will be held before the start of school (School Year 2008-2009). Senior Abt Associates staff that have developed the classroom observation instrument and have experience using similar kinds of protocols in other studies will conduct the trainings. The training sessions will be held in a location that is most convenient and cost-efficient, given the location of the data collectors and training staff. A training manual will be prepared in advance of the training session. The manual will include the instruments and the procedures and protocols to follow for conducting the data collection. The training will combine an item-by-item discussion of the instrument and protocol, followed by hands-on practice time (e.g., practice observations of videotaped literacy lessons). In addition, classroom observers will be trained on the protocol for collecting a 20-minute audiotape recording of a literacy lesson. Data collectors will also need to establish reliability with the trainers on the observation instrument.

Prior to the start of the school year, Abt Associates staff will train SERVE data collectors to conduct the classroom observation measure. The training meeting will span three days, involving (1) a detailed discussion of the procedures for administering the instrument and of the specific items and (2) discussions of how to code videotaped samples of kindergarten literacy instruction. In addition, SERVE data collectors will independently code additional videotaped samples of kindergarten literacy instruction until they achieve 85% reliability with expert raters from Abt Associates.

Abt Associates staff will provide classroom data collectors with the same general training provided to student assessors, described above, for working with schools and teachers and for maintaining privacy and confidentiality. Abt Associates staff will also develop criteria prior to training to establish that data collectors meet standards required for data collection. Staff from SERVE will closely oversee data collectors in the early weeks to identify and correct any problems early.

Procedures for collecting classroom data. Classroom observations will be conducted in both treatment and control classrooms for the full block of time devoted to literacy instruction during the school day. Trained data collectors, hired and overseen by SERVE staff, will contact each teacher in advance of the classroom observation in order to schedule the visit on a day that is convenient for the teacher. At this time, the SERVE data collector will explicitly indicate that s/he wants to observe the period in the day spent on literacy instruction and will ask the teacher what time literacy instruction will occur. The data collector will visit the classroom and conduct the classroom observation at the time indicated by the teacher. If a change in the teachers' schedule occurs unexpectedly causing the data collector to miss the period of literacy instruction, the classroom observation will be rescheduled. Data collectors will be responsible for working with the schools and teachers to identify appropriate times for visits and to follow established procedures (e.g., signing in at the front desk, wearing an ID badge).

As mentioned above, the classroom observation and the audio recording of the literacy lesson will be conducted two times – in the fall (2008) and the spring (2009) of the intervention year. Prior to the start of data collection, consent forms from teachers and parents will have been obtained. Data collectors will follow the protocol established for observing the classroom. Data collectors will sit in the back of the classroom during the observations and will be as unobtrusive as possible. At the same time as each classroom observation, data collectors will also collect a 20-minute audio recording of a literacy lesson in the classroom.

Student demographic information from extant sources. Staff from Empirical Education Inc. will collect extant demographic information about students from school administrative records, including information regarding age, gender, race, ethnicity, eligibility for free or reduced-price school meals, special education status, and status as an English-language learner. The data collection form for gathering extant data is attached in Appendix L. Child demographic information will come directly from electronic files. Empirical Education will remove identifying information (i.e., children's names) but will include the same assigned ID number used for the child assessments, so that the demographic and assessment data files can be merged into a single electronic database.

2b. Stratification and sample selection

Sample selection, stratification of schools based on prior experience with reading initiatives, further stratification based on observable school characteristics, and random assignment within strata have been described above in response to Question B.1. Schools that have had experience with similar reading initiatives will be grouped together, as will schools that have not had previous experience with other reading initiatives. Teachers will be selected from each participating school in Spring 2008, prior to random assignment. Schools within each stratum then will be randomly assigned in Spring 2008 to either the PAVE treatment or the control condition. Students from each participating classroom will be selected in Fall 2009, after random assignment and after teachers in the treatment group have received the PAVE training.

2c. Estimation procedures

Impacts of PAVE both on students and on teachers will be estimated as described below.

Immediate impacts on students. To examine the impact of PAVE on students at the end of kindergarten, we will use a hierarchical linear model (HLM), which provides us with an estimate of the average impact of the intervention on children across all schools at a given time point (e.g., at the end of the kindergarten year). The HLM is particularly appropriate for this evaluation since we have a multilevel design with students nested within classrooms and schools[3]. Students in our evaluation are clustered within schools, and the treatment occurs for all kindergarten students within a school. HLM enables us to adjust standard error estimates to account for the nesting of students in classrooms. This adjustment is crucial, as the unit of random assignment is the school not the student. In addition, HLM allows us to determine what proportion of the total variation in student outcomes occurs at the school level and what proportion occurs at the student-level. HLM also allows us to use both school-level and student-level covariates to account for variation in student outcomes.

We can write the model in hierarchical form. The level-1, or student-level model is:

$$Y_{ij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_q (X_{qij} - \overline{X}_q) + \varepsilon_{ij} \qquad (1)$$

where,

[3] There is no classroom-level equation in the model due to our earlier assertion that we do not expect much variance in student achievement between kindergarten classrooms within a given school.

Y_{ij} = an outcome measure (e.g., PPVT score) of the i^{th} student in the j^{th} school;

$X_{qij} - \overline{X}_q$ = a vector of baseline student characteristics, centered around the overall mean, across the *i* students in the *j* schools;

β_{0j} = the covariate adjusted mean value of the outcome measure for school *j;*

β_q is a vector of *q* regression coefficients, which capture the effects of the X_{Qij} predictor variables on outcomes averaged across all students at all schools; and

ε_{ij} = the student-level residual or error term of the i^{th} student in the j^{th} school (i.e., the deviance between the i^{th} student in school *j* and the average student in school *j*, controlling for student-level covariates). The assumed distribution of these residuals is normal, with mean = 0, and variance $=\sigma^2$.

The level-2 or school-level model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j - \overline{T}) + \gamma_{02}(READ_j) + \sum_{k=1}^{K} \gamma_{0k}(Z_{kj} - \overline{Z}_k) + \upsilon_{0j} \qquad (2)$$

where,

γ_{00} = the grand or overall mean value of the outcome measure across schools;

γ_{01} = the average treatment effect on school mean student scores (i.e., the mean difference between treatment and control schools);

$T_j - \overline{T}$ = treatment status, where T = 1 for a school assigned to the PAVE treatment and T = 0 for a school assigned to the control group, centered at the mean value of T across schools (i.e., $\overline{T}$);

γ_{02} = the average effect of prior experience with a reading initiative on school mean student scores (i.e., the mean difference between schools with a prior reading initiative and schools with no prior reading initiative);

$READ_j$ = an indicator variable for school's prior experience with a reading initiative, where READ = 1 for a school with a prior reading initiative and READ = 0 for a school with no prior reading initiative;

γ_{0k} = a vector of *k* regression coefficients indicating the effect of each school-level covariate on school mean scores;

$Z_{kj} - \overline{Z}_k$ = a vector of *k* school-level covariates (*k* equals the number of school-level covariates), each centered around the grand mean; and

υ_{0j} = the error term for the j^{th} school (i.e., the deviance between the j^{th} school and the grand mean, controlling for school-level covariates and treatment effects). The distribution is assumed to be normal, with mean = 0, and variance $=\tau^2$.

The estimate of γ_{01} indicates whether there is a significant impact of the PAVE treatment on the specified student outcome. A positive and statistically significant estimate of γ_{01} indicates that the PAVE intervention does impact student vocabulary (or broader literacy) outcomes. The magnitude of γ_{01} indicates the estimated magnitude of the impact, i.e., participation in PAVE is associated with an estimated γ_{01} point difference in scores (or standard deviation difference, depending on the units of measurement for γ_{01}) of students in PAVE schools compared to students in non-treatment schools.

In addition to examining the impact of PAVE on students overall, we will examine its impact on subgroups of students, such as boys and girls. By analyzing subgroups of students, we can determine if there are differential effects of the PAVE intervention for certain subsets of students. Specifically we may be interested in knowing whether the effects of PAVE systematically differ for boys and girls. To address this question we will take one of two approaches.

The first approach would be to include an interaction of the treatment effect with a subgroup variable in the HLM model (e.g., treatment*BOY, where BOY = 1 for boys and 0 for girls). The estimated parameter of such an interaction would indicate if there were additional effects of PAVE for boys or girls. This is a conventional approach; however, it would require the assumption that the variance in outcome scores is constant for boys and girls and that gender is not correlated with the marginal effects of the other covariates in the model, which may or may not be a tenable assumption. Specifically, testing an interaction between a level-1 student characteristic and the level-2 treatment predictor is represented in a combined model[4], which is generated by substituting the level-2 equation for β_{0j} into the level-1 equation. Before substituting the level-2 equation into the level-1 equation, we can also express the level-2 equation for β_{1j} as follows: $\beta_{1j} = \gamma_{10} + \gamma_{11}(T_j - \bar{T}) + \upsilon_{1j}$, and substitute this equation into the level-1 equation as well. The combined model, represented below, allows the treatment effect for boys to vary from the treatment effect for girls:

$$Y_{ij} = \gamma_{00} + \gamma_{01}(T_j - \bar{T}) + \gamma_{02}(READ_j) + \gamma_{10}BOY_{ij} + \gamma_{11}(BOY_{ij} * [T_j - \bar{T}]) + \varepsilon_{ij} + \upsilon_{0j} + \upsilon_{1j} \qquad (3)$$

In this equation, γ_{01} indicates the average impact of the treatment for girls, and γ_{11} indicates the differential in the average impact of PAVE for boys compared to girls.

The second approach would be to break the entire sample into subgroups, (e.g., one sample entirely comprised of boys, the other of girls), and estimate the treatment effect for each of the two samples. Once these impacts are estimated, we can compare the means and variances of the estimates of the subgroups to determine if there are differential impacts between these two groups. Differences in impact between the two groups (i.e., the interaction effect) will then be calculated via a t-test. This is also a sensible approach; however, the smaller sample sizes for each subgroup inherently make impact estimates for each group less precise.

Measuring changes in impacts on students over time. With the addition of another school year of data (i.e., first grade), we will be able to extend the cross-sectional model to look at changes in students over time, using longitudinal linear growth modeling. A hierarchical linear approach, such as the one described below, is a convenient tool for modeling patterns of change among children over time. The model is hierarchical in the sense that multiple observations are nested within individual students who are, in turn, nested within classroom/schools. For purposes of illustration, we present here a three-level hierarchical linear growth model. The first level of our model represents each student's development in the form of an individual linear growth trajectory, whose parameters then become the outcome variables in the between-student level of

[4] Student-level X covariates, other than BOY, and school-level Z covariates are not shown in this model to facilitate the illustration of the interaction between the level-1 covariate, BOY, and the level-2 treatment indicator.

our model. Those parameters can vary among students and schools as a function of child- or teacher/school-level variables. The within-student or repeated observations model (Level-1) is denoted as follows:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\left(G_{tij} - C\right) + \alpha_{tij} \qquad (4)$$

where,

Y_{tij} = an observed outcome measure (PPVT score) for student *i* in school *j* at time *t*;

G_{tij} = the grade for student *i* in school *j* at time *t;*

C = the centering parameter set to a particular grade level (e.g., kindergarten)*;*[5]

π_{0ij} = the score (intercept) for student *i* in school *j*, defined in grade *C;*

π_{1ij} = the linear growth rate parameter (rate of change from one grade to the next) for student *i* in school *j;* and

α_{tij} = a random error term for student *i* in school *j* at time *t*. The distribution of within-student errors is assumed to be normal, with mean=0 and variance=ϕ^2.

In the between-student model (Level-2), variation in the growth parameters π_{ij} can be modeled as a function of student background characteristics (e.g., sex, ethnicity, free/reduced school lunch eligibility). In this second level of the model, the π_{ij} are random outcome variables. A between-student model can be formulated for both the intercept, π_{0ij}, and linear growth rate parameter, π_{1ij}, as follows:

$$\pi_{0ij} = \beta_{00j} + \sum_{q=1}^{Q} \beta_{0q}(X_{qij} - \overline{X}_q) + \delta_{0ij} \qquad (5)$$

$$\pi_{1ij} = \beta_{10j} + \sum_{k=1}^{K} \beta_{1k}(X_{kij} - \overline{X}_k) + \delta_{1ij} \qquad (6)$$

where,

π_{0ij} = the score for student *i* in school *j*, defined in grade *C* (i.e., the level-1 intercept);

π_{1ij} = the linear growth rate parameter (from level-1 model) for student *i* in school j;

β_{00j} = the average covariate adjusted score across all students in school *j* in grade C,

β_{10j} = the average covariate adjusted linear growth rate of outcomes across all students in school *j*;

β_{0q} = a vector of *q* = 1...*Q* regression coefficients, which capture the effects of the X_i predictor variables on the student-specific outcome score in grade C averaged across all students at all schools;

β_{1k} = a vector of *k* = 1...*K* regression coefficients, which capture the effects of the X_i predictor variables on the student-specific linear growth parameter across all students at all schools;

$X_{qij} - \overline{X}_q$ and $X_{kij} - \overline{X}_k$ = measured background characteristics for student *i* in school *j*, each centered at the overall mean value, which may differ when predicting the average student outcome score in grade C and the linear growth rate in outcomes; and

[5] In this model, the centering parameter, *C*, is chosen to be a meaningful point in time, so that π_{0ij} is made interpretable, representing the student's ability at that grade level (Bryk & Raudenbush, 1992). For example, given the range of grade levels in the evaluation, we can set *C* at either kindergarten or 2nd grade) depending on the type of inference desired.

δ_{0ij} = random error term indicating the deviance between the score for student *i* in school *j* in grade C and the average score for students in school *j* in grade C, after controlling for the vector of *Q* student-level characteristics;

δ_{1ij} = random error term indicating the deviance between the linear growth rate for student *i* in school *j* and the average linear growth rate for students in school *j*, after controlling for the vector of *K* student-level characteristics.

The between-student error terms are assumed to have a bivariate normal distribution, with mean 0, variances σ_0^2, σ_1^2 and covariance $\sigma_{10} = \sigma_{01}$. These assumptions are expressed as follows:

$$\begin{bmatrix} \delta_{0ij} \\ \delta_{1ij} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right).$$

The impact of the PAVE treatment on student growth is tested in a school-level model. In the school-level model (Level-3), variation in school mean scores of children in grade C (*β_{00j}*) and variation in school mean linear growth rates (*β_{10j}*) can be modeled as a function of PAVE treatment status and other school characteristics. The school-level model is specified as follows:

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(T_j - \bar{T}) + \gamma_{002}(READ_j) + \sum_{q=1}^{Q} \gamma_{00q}(Z_{qj} - \bar{Z}_q) + \upsilon_{00j} \qquad (7)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}(T_j - \bar{T}) + \gamma_{102}(READ_j) + \sum_{k=1}^{K} \gamma_{10k}(Z_{kj} - \bar{Z}_k) + \upsilon_{10j} \qquad (8)$$

where,

β_{00j} = the average covariate-adjusted score across all students in school *j* in grade C;

β_{10j} = the average covariate-adjusted linear growth rate across all students in school *j*;

γ_{000} = the grand or overall covariate-adjusted mean value of student score in grade C across all students in all schools;

γ_{100} = the grand or overall covariate-adjusted mean value of linear growth rate across schools;

γ_{001} = the average treatment effect on school mean student score in grade C;

γ_{101} = the average treatment effect on school mean linear growth rate;

$T_j - \bar{T}$ = treatment status, where T = 1 for a school assigned to the PAVE treatment and T = 0 for a school assigned to the control group, centered at the mean value of T across schools;

γ_{002} = the average effect of prior experience with a reading initiative on school mean student score in grade C;

γ_{102} = the average effect of prior experience with a reading initiative on school mean linear growth rate;

$READ_j$ = an indicator variable for school's prior experience with a reading initiative, where READ = 1 for a school with a prior reading initiative and READ = 0 for a school with no prior reading initiative;

γ_{00Q} = a vector of *Q* regression coefficients indicating the effect of each school characteristic on school mean student score in grade C;

$Z_{Qj} - \overline{Z}_Q$ and $Z_{Kj} - \overline{Z}_K$ = school covariates for school *j*, each centered around the grand mean, which may differ when predicting the average student outcome score in grade C at school *j* and when predicting the average linear growth rate of student scores at school *j*

γ_{10K} = a vector of *K* regression coefficients indicating the effect of each school characteristic on school mean linear growth rate; and

υ_{00j} = the error term for the *j*th school's status (i.e., the deviance between mean student score for the *j*th school's mean outcome score and the grand mean student score, controlling for school-level covariates and treatment effects); and

υ_{10j} = the error term for the *j*th school's linear growth rate (i.e., the deviance between mean linear growth rate for the *j*th school and the grand mean linear growth rate, controlling for school-level covariates and treatment effects).

The school-level error terms are assumed to have a bivariate normal distribution, with mean 0, variances τ_0^2, τ_1^2 and covariance $\tau_{10} = \tau_{01}$. These assumptions are expressed as follows:

$$\begin{bmatrix} \upsilon_{00j} \\ \upsilon_{10j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{bmatrix}\right).$$

The estimates of γ_{101} and γ_{001} indicate whether there is a significant impact on a specified student outcome of the PAVE treatment on average linear growth and on average student score at grade C, respectively. A positive and statistically significant value of γ_{101} indicates that the PAVE intervention does impact average growth in student vocabulary (or broader literacy) outcomes. A positive and statistically significant value of γ_{001} indicates that the PAVE intervention does impact average grade C score in student vocabulary (or broader literacy) outcomes. The magnitude of γ_{101} indicates the estimated magnitude of the impact, i.e., participation in PAVE is associated with an estimated γ_{101} point difference in average growth per grade (or standard deviation difference in average growth per grade, depending on the units of measurement for γ_{101}) of students in PAVE schools compared to students in non-treatment schools. Similarly, the magnitude of γ_{001} indicates the magnitude of the PAVE impact on school average student score in grade C.

Impacts on teachers. The impact of the PAVE intervention on teacher and classroom practices, controlling for teacher and school characteristics, will be estimated using a multilevel model, in order to account for the clustering of two teachers per school. The model will include a teacher-level (Level-1) and a school-level (Level-2). Because of the limited degrees of freedom at Level-1 (due to sampling only two teachers per school), we will control for teacher characteristics at the school-level. For each teacher characteristic, we will calculate the average value for the school.

The multilevel model will be of the following form[6]:

[6] For teacher outcomes that are not continuous (i.e., categorical or dichotomous) we will apply a similar analytic strategy with a model appropriate for the outcome. For a dichotomous outcome, we will specify a logistic regression model, estimating log-odds as a function of participation in the intervention, controlling for teacher and school

Level-1: Teacher-level model:

$$\mathrm{Y}_{ij} = \beta_{0j} + \varepsilon_{\mathrm{ij}} \tag{9}$$

where,

Y_{ij} = an outcome *Y* for the i^{th} teacher in the j^{th} school;
β_{0j} = an estimate of the mean value of the outcome Y for school *j*;
ε_{ij} = the residual for the i^{th} teacher in the j^{th} school. The level-1 residuals are assumed to be normally distributed with mean=0 and variance = σ^2.

Level 2: School-level model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j - \overline{T}) + \gamma_{02}(READ_j) + \sum_{q=1}^{Q}\gamma_{0q}(X_{qj} - \overline{X}_q) + \sum_{k=1}^{K}\gamma_{0k}(Z_{kj} - \overline{Z}_k) + \upsilon_{0j} \tag{10}$$

where,

γ_{00} = the grand mean outcome for all schools;
γ_{01} = the average treatment effect on the school mean teacher outcome;
$T_j - \overline{T}$ = treatment status, where T = 1 for a school assigned to the PAVE treatment and T = 0 for a school assigned to the control group, centered at the mean value of T across schools;
γ_{02} = the average effect of prior experience with a reading initiative on school mean teacher outcome (i.e., the mean difference between schools with a prior reading initiative and schools with no prior reading initiative);
$READ_j$ = an indicator variable for school's prior experience with a reading initiative, where READ = 1 for a school with a prior reading initiative and READ = 0 for a school with no prior reading initiative;
γ_{0q} = a vector of regression coefficients indicating the effects of *Q* teacher characteristics, averaged for school *j*;
$X_{Qj} - \overline{X}_Q$ = a vector of Q teacher characteristics, averaged for school *j*, and centered at the grand mean;
γ_{0k} = a vector of regression coefficients indicating the effects of *K* school characteristics;
$Z_{Kj} - \overline{Z}_K$ = a vector of school characteristics for school *j*, centered at the grand mean; and
υ_{0j} = the level-2 random effects, or school *j*'s deviation from the grand mean of the teacher outcome, not accounted for by treatment status or control variables. We assume that the school-level error terms are normally distributed with mean=0 and variance=τ^2.

Our estimate of γ_{01} indicates whether there is any significant impact of the PAVE treatment on a specified teacher outcome. Thus, a positive and significant estimate of γ_{01} indicates that the PAVE intervention does influence how teachers conduct classroom activity and implement program features.

characteristics. For a categorical outcome, we will specify a multinomial logit model, estimating the log-odds for each category of the outcome as a function of participation in the intervention, controlling for teacher and school characteristics.

2d. Degree of accuracy needed

A statistical power analysis was conducted to determine the appropriate sample size for this study. Expected effect sizes used in the power analysis were based on previous research on PAVE. Quasi-experimental research indicates that the expressive and receptive language scores on standardized assessment measures are between .20 and .43 standard deviations higher for children in PAVE classrooms than in comparison classrooms. However, meta-analyses that have compared average effect sizes of educational interventions from randomized versus quasi-experimental designs suggest that, on average, quasi-experimental designs produce larger estimates of effects. Taking this into account we plan to recruit a sample that will allow us to detect effect sizes at the lower end of this range. The following assumptions were used in the power analysis:

(a) The design is hierarchical, with students nested within classrooms, which in turn are nested within schools, and variation in student achievement between classrooms should be minimal;
(b) The intra-class correlation at the school level will range between 0.05 and 0.15 because of the project's focus on recruiting schools exclusively within high-poverty communities;
(c) The overall correlation at the individual level between standardized pretests and posttests in kindergarten will be 0.70 (i.e., student-level $R^2 = .50$);
(d) The correlation between any other student-level covariate and the outcome measure is .00, which does not raise the student-level R^2;
(e) The school-level R^2 is .00 (i.e., the correlations between any school level covariates and the outcome measure are all .00);[7]
(f) By the end of the study, the attrition rate will equal approximately 20%;
(g) A two-tailed test of significance will be conducted at the 0.05 level; and
(h) The power to detect effects is .80.

The following equation was used to estimate minimum detectable effect sizes (MDES) for 60-80 schools (*J*) and ICCs (ρ) ranging between .05 and .10.

$$MDES(\hat{\beta}_0) = \sqrt{1+(20-1)\rho} * \sqrt{\frac{\rho}{(.5*.5)J} + \frac{(1-\rho)(1-.25)}{(.5*.5)20}}$$

Having a sample size between 60 to 80 schools should give us acceptable power to detect effect sizes hypothesized for this study. Table B.4 presents results of our power calculations for this range. Based on the power calculations, we will recruit a sample of at least 60 schools. With two classrooms per school and 10 students per classroom and a conservative ICC of .15, we will need to recruit 60 schools to detect an effect size of 0.27 standard deviations at the student level, which seems reasonable given results from previous, quasi-experimental studies.

[7] Assuming a higher school-level R^2 or additional explanatory power for school-level covariates (other than for pretest) would actually increase our power estimates. It is possible that the school-level and/or student-level R^2 could be higher than we assume; however, we are erring on the side of being more conservative in our power calculation to ensure that we will have sufficient power to detect hypothesized effects.

Because the sample will necessarily be restricted to schools that are able and willing to participate, bias may be introduced into the sample, which could either overestimate or underestimate intervention effects. The selection bias may favor intervention effects if the willing and able schools are more capable of effectively implementing PAVE than schools that are not willing and able to participate or have students who can benefit more from a given level of implementation. Alternatively, schools that are performing most poorly may be more eager for improvement than better-performing schools and as a result may participate at a higher rate. In this situation, intervention effects might be underestimated if poorer performing schools are less capable of implementing PAVE or have students who cannot benefit as much from a given level of implementation.

Table B.4
Power Analysis Summary Table for Student Outcomes: Minimum Detectable Effect Size by Number of Schools and Assumed Intra-class Correlation
(two-tailed test at .05; n = 20 students per school; attrition =. 20; $R^2 = .50$; power = .80).

Intra-class Correlation	60 schools 1200 students	70 schools 1400 students	80 schools 1600 students
.05	.22	.20	.19
.10	.24	.23	.21
.15	.27	.25	.23

The study is designed with the goal of detecting effects on students; consequently, we will not have the same level of power for detecting effects on teachers. Nonetheless, we also want to compare PAVE teachers with control teachers to see if they have better instructional behaviors, a necessary precursor in our conceptual model to observing effects on students. Our power for these analyses will be much more limited, given a sample of two teachers per school.

The equation used to estimate minimum detectable effect sizes (MDES) for impacts on teachers, given 60-80 schools (*J*) and an ICC (ρ) ranging between .05 and .15 is:

$$MDES(\hat{\beta}_0) = \sqrt{1+(2-1)\rho} * \sqrt{\frac{\rho}{(.5*.5)J} + \frac{(1-\rho)}{(.5*.5)2}}$$

Table B.5 shows the results of the power analysis for detecting effects on teachers. If we recruit 60 schools or more, we will be able to detect an effect size at the teacher level of at least 0.56 standard deviations. This minimum detectable effect size is appreciably greater than the one for detecting differences at the student level, but it is still appropriate given the expected impact of the intervention on teachers’ instructional practices. Because the intervention is intended to directly impact teachers’ instructional practices, we expect larger effects on teachers than we expect on students. Furthermore, we expect that impacts on teachers must be on this order of magnitude in order to translate into impacts on students.

Table B.5
Power Analysis Summary Table for Teacher Outcomes: Minimum Detectable Effect Size by Number of Schools
(two-tailed test at .05; intra-class correlation = .15; n = 2 teachers per school; power = .80).

60 schools 120 teachers	70 schools 140 teachers	80 schools 160 teachers
.56	.51	.48

2e. Unusual problems requiring specialized sampling procedures

We do not anticipate any unusual problems requiring specialized sampling procedures.

2f. Use of periodic data collection cycles to reduce burden

During the intervention year (School Year 2008-2009), pretest data on teachers and classrooms in both the treatment and control groups will be collected in Fall 2008. Posttest data on teachers and classrooms will be collected in Spring 2009. Follow-up data on treatment group teachers and classrooms will also be collected in the fall of the subsequent school year (Fall 2009).

Data on one cohort of students will be collected at the beginning of the kindergarten year (Fall 2008) as a pretest, at the end of the kindergarten year (Spring 2009) as a posttest, and at the end of the first grade year (Spring 2010) as a follow up.

3. Describe the methods to maximize response rates and to deal with issues of non-response. The accuracy and reliability of information collected must be shown to be adequate for intended uses. For collections based on sampling, a special justification must be provided for any collection that will not yield "reliable" data that can be generalized to the universe studied.

We have not begun recruitment efforts at the district, school, teacher, and parent/student levels yet. At the end of April there was an initial meeting between the REL-Southeast Director, Vocab Study Manager, and key administrators in MS Department of Education, including the Superintendent of Education, Dr. Hank Bounds; the Executive to the Superintendent for Instructional Programs and Services, Beth Sewell; and the Director of the Office of Reading, Early Childhood, and Language Arts, Robin Miles.

Our next step is for project staff to work with the Mississippi State Superintendent's office to determine the most efficient and effective way to recruit school districts and schools for this study. The State School Superintendent has expressed support for this study in recent conversations with the REL-Southeast Director and has agreed to help in the recruitment effort. We have scheduled a kick-off meeting for September 17, 2007 for the PAVE intervention developers and the Evaluation Project Director to make presentations to these key state-level personnel and answer questions about the intervention and the study. We are anticipating that careful groundwork with the Superintendent's office will help us to achieve high response rates and to maintain schools in the sample. We will focus our recruitment efforts at the district level, with the expectation that this will lead to greater participation rates for all elementary schools within a district. To secure buy-in at the district level, we will speak with all the necessary and appropriate individuals, including district superintendents and/or assistant superintendents, and provide them with a letter of support from the State Superintendent (see Appendix M), a description of the intervention (see Appendix N), and a study overview with a clear description of the plan for the random assignment of schools, including plans for providing PAVE professional development to teachers in control schools during the summer following the intervention year (see Appendix O). We will also have senior staff available to answer their questions or provide additional information. We will also prepare materials as necessary for district IRB or other approval processes. We will ask each district superintendent to sign a partnership agreement that conveys the responsibilities of the REL-SE study staff and the school district (see Appendix P). Once districts have agreed to participate, we will work with them to enlist support and participation from the schools and teachers at the targeted elementary schools. (Additional recruitment materials, including a description of the intervention aimed at teachers and recruitment letters for principals and teachers are included in Appendices Q, R, and S, respectively.) If a school agrees to participate, we ask the principal to sign a school partnership agreement (see Appendix T).

We will work with school administrators to optimize our recruitment efforts and help us achieve high consent rates among teachers. Teachers will be provided with intervention materials for their classrooms, including books and curriculum units with intervention-related activities, which University of Georgia estimates to be worth about $750. The materials will be provided to treatment teachers for the intervention year, and control teachers, along with PAVE training, at the end of the intervention year. In an effort to offer teachers further incentive to participate, our plan is to talk with members of the Mississippi State Department of Education about what can be offered to teachers as an incentive to participate, including professional development credits.

In our efforts to obtain parental permission, we will work with the State of Mississippi and participating school districts to determine if they have any requirements we will need to follow for this process. Our experience shows that different districts often handle the issue differently, and sometimes schools will have unique procedures that they follow. We will encourage participating schools to assign someone on their administrative staff to serve as a study liaison. We would work with this person to get the parental permission forms (which will have been approved by the Institutional Review Boards at the University of North Carolina at Greensboro, University of Georgia, and Abt Associates Inc.) to parents as soon as possible, and to follow up with those parents not responding. Abt Associates has had fairly good success using this approach in the past. We plan to enclose parental consent forms in the packet of advance materials sent to home prior to the start of the school year or on the first day of school, which contain information about teacher assignment, school procedures, and numerous school forms (e.g., emergency contacts, immunization history). The liaison would collect the returned consents and follow up persistently with those parents who have not yet returned the consent forms.

Critical to our strategy for obtaining high response rates from the recruited sample are data collectors who: work with schools and teachers to schedule visits for optimal data collection times (when the children are alert and when the session will not be interrupted); are skilled at building rapport with children; and work thoroughly but efficiently through each data collection step. We will explore providing children with a small token after each data collection session to provide additional motivation to complete the assessment session.

Attrition will be carefully monitored in this study. As discussed above, our power analysis assumes a 20% attrition rate in children from the start of the study through the end of first grade. Despite this conservative assumption, we expect to have upwards of 85-90% of the sample intact at the end of kindergarten and first grades. Our expectations about the response rate are based in part on the fact that we will be sampling from a pool of students with parental permission. Although some parents can be expected to refuse to allow their children to participate, such refusals will not lower the response rate among children selected for the sample, as parental permission will have been obtained prior to sample selection. We expect that most children will comply with child assessments. We anticipate that some students will refuse to participate and that we will be unable to schedule assessments with some students due to absences or leaving the school. Nonetheless,we assume that mobility among students at this age is fairly low. Despite the fact that mobility can be high among low-income households, this movement is not necessarily between schools. We expect that student non-compliance, absence, and mobility will contribute to a loss of less than 15% of the sample of students by the first grade follow-up one year after the intervention.

Teacher attrition is expected to be negligible during the first year because of contractual obligations. All data examining impacts of PAVE on teaches will be collected during the intervention year. Following the intervention year, we will collect data only from intervention teachers, not from control teachers. At the end of the intervention year, we will collect information regarding treatment teachers' expected placement for the next study year. If the teacher moves to a kindergarten classroom at another study school, we can continue to include them in the study. Beyond the intervention year, we expect teacher attrition to be high. For that reason, we will not collect data on a second cohort of students. According to the Schools and Staffing Survey, in 2004-2005, the annual rate of teacher turnover nationally among public school teachers was 16.5% (Marvel, Lyter, Peltola, Strizek, and Morton, 2007). The turnover rate is estimated to be higher, hovering around 20%, among early childhood and elementary school teachers, among teachers in high poverty schools, or among teachers in schools with a high percentage of minority students. The lower response rate during the year following the intervention will affect our investigation of the sustainability of the PAVE intervention among treatment teachers only, but not our examination of PAVE impacts.

We have proposed providing monetary incentives, if approved by the MS Department of Education, to teachers in the control group to compensate them for the time required to attend the PAVE training and other meetings outside the school day, as described in Section A9. Specifically, teachers would receive $200 for attending the PAVE training during the summer and $150 after completion of all the after-school peer discussions. Teachers in the control group are not required to devote time to attend training or meetings and thus will not receive monetary compensation.

We do not believe that providing treatment teachers with incentives for demands on their time will make them more cooperative than control teachers about scheduling observations and assessments. Although it is possible that treatment teachers could be more cooperative, a counter-argument could be made that some treatment teachers could be less cooperative than control teachers. Treatment teachers are expected to alter their instructional practices, while control teachers are not. If treatment teachers do not adopt the PAVE intervention practices, they may be less likely to cooperate with scheduling observations and assessments than control teachers who may be comfortable having their routine teaching practices observed. Similarly, control teachers who want the PAVE intervention may be more motivated to remain in the study than treatment teachers who have already received the PAVE training. Because we can hypothesize scenarios in which treatment teachers would be more cooperative and scenarios in which control teachers would be more cooperative, we conclude that neither group of teachers can be assumed to be more cooperative than the other.

Incentives for completion of the teacher demographic questionnaire and the implementation challenges interview will not be provided to teachers in either the treatment or control group. The demographic questionnaire will be distributed to and collected from teachers during a training session or a classroom visit. Teachers will be asked to submit the completed questionnaire either before leaving the training session or before the data collector leaves the classroom. Collecting questionnaires on-the-spot and in person will help encourage teachers to complete them. Similarly, the implementation challenges interview will be completed in person after a fidelity assessment visit. Conducting the interview individually and in person when data collectors are already in the classroom for another purpose will facilitate a high response rate.

We will assess the extent of bias resulting from sample attrition. We will document the response rate in treatment and control groups for both teachers and students, paying particular attention to whether the response rates are unequal and if so what might account for unequal attrition. If we find evidence of unequal attrition rates, we will adjust for that bias accordingly using statistical imputation and weight adjustments.

4. ***Describe any tests of procedures or methods to be undertaken. Testing is encouraged as an effective means of refining collections of information to minimize burden and improve utility. Tests must be approved if the call for answers to identical questions from 10 or more respondents. A proposed test or set of tests may be submitted for approval separately or in combination with the main collection of information.***

We conducted a two-day pilot test of the data collection procedures and instruments at one elementary school in a rural Georgia town, approximately 50 miles northeast of Atlanta. The town has a population of 10,201 and an estimated median household income in 2005 of $34,600. The students in the elementary school are 68.9% White, 18.7% African American, 7.8% Latino, and 4% Asian. The demographic characteristics of the pilot site differ from the Mississippi Delta, as the population in the pilot site is predominantly White rather than African American and has a higher average household income than in the Mississippi Delta.

Table B.6 indicates the pilot test schedule, the number of respondents completing assessments and questionnaires, and the number of classrooms that were observed using standard observational measures. On the first day of the pilot test, kindergarten teachers participated in the PAVE professional development training, at which time they completed the Teacher Demographic Questionnaire. On the same day, the child assessment battery was administered to

6 students. On the second day, we conducted classroom observations: (1) in three classrooms, we tested the classroom observation instrument developed for this study; and (2) in two classrooms, we tested the Fidelity Rating Scale that will be used to examine whether PAVE practices are sustained into the year following the intervention.

Table B.6
Pilot Test Schedule, Number of Respondents, and Average Administration Time

	Number of Respondents	Average Administration Time
Day 1		
Teacher Demographic Questionnaire	6 Teachers	10 minutes
Child Assessment Battery:	6 Students	
PPVT-4		15 minutes
EVT-2		15 minutes
Elicited Language Sample		13 minutes
Day 2		
PAVE Fidelity Rating Scale	2 Classrooms	1 hour
Classroom Observation Measure	3 Classrooms	1 hour

Teacher Demographics Questionnaire

The 6 kindergarten teachers completed the Teacher Demographics Questionnaire. Based on this pilot administration, we have modified questions on the form and created an alternative version of the questionnaire for the kindergarten paraprofessionals (i.e., Paraprofessional Demographics Questionnaire). Questions on the Teacher Questionnaire about formal education and alternative certification are now distinct instead of combined. The Paraprofessional Questionnaire asks for information about special training and relevant experiences, as well as formal education and certification. (See Appendices E and F for the demographics questionnaires.)

Child Assessments

While teachers attended the PAVE training, experienced University of Georgia data collectors administered the kindergarten child assessments to six kindergarten students in the school. These students were selected at random from students with signed parental permission to participate. The child assessment measures used in this study have been used in previous studies, but the goal of including them in the pilot was to obtain information about data collection procedures and administration time. Staff from Abt Associates observed the child assessments, documenting the start and end time for each assessment and taking notes on how procedures worked. The average administration times were consistent with prior expectations. On average, the child assessment battery was 42.8 minutes, ranging from 33 to 59 minutes. The average administration time for the Peabody Picture Vocabulary Test was 15 minutes, ranging from 12 to 18 minutes. The average administration time for the Expressive Vocabulary Test was 15 minutes, ranging from 7 to 16 minutes for 5 students and lasting 24 minutes for a student with a speech problem. For the elicited language sample (i.e., measurement of lexical diversity), administration time averaged 13 minutes, ranging from 8 to 18 minutes.

From the pilot test, we confirmed that the order of administration of the child assessments was effective. The administration manual for the PPVT and EVT indicates that the PPVT should be administered first, followed by the EVT. We followed this procedure in our test, and will do so in the full study. In addition, we confirmed that eliciting the student language sample after completing the standardized assessments is the best approach, as students have become comfortable enough with the tester by that point to produce a sufficient language sample for measuring lexical diversity.

Classroom Observations

Day two of the pilot test involved classroom observations, during which we tested two observation procedures and instruments: (1) a fidelity rating scale (see Appendix J) and (2) a classroom observation of literacy teaching practices (see Appendix G). In order to maximize the number of classrooms observed and the number of observation instruments tested during one day, we conducted each observation for one hour rather than the two to three hours planned for the full study.

PAVE Fidelity Rating Scale

Three observers from Abt Associates tested a newly developed instrument for measuring teachers' fidelity to the PAVE instructional model in two classrooms. The PAVE Fidelity Rating Scale involves on-the-spot ratings of teachers' and paraprofessionals' adherence to PAVE instructional components on a continuum from low to high implementation. The instrument will not be used in this study until Fall 2009. We are continuing to refine the instrument based on our experiences with it in the pilot test. In addition, we are consulting with PAVE intervention designers regarding questions and clarifications about implementation fidelity and are adjusting the assessment accordingly.

Classroom Observation of Literacy Teaching Practices

Observers from Abt Associates tested a newly developed instrument for measuring teachers' vocabulary and broader literacy instructional practices in three classrooms. Based on the pilot test, a number of modifications to the instrument were identified, including changes in coding categories and instrument format. In addition to changing the observation instrument, the pilot identified some items to be included in the coding manual, including more specific definitions of coding categories, and a discussion of procedures for ensuring greater consistency across raters in timing observational intervals.

5. ***Provide the name and telephone number of individuals consulted on statistical aspects of the design and the name of the agency unit, contractor(s), grantee(s), or other person(s) who will actually collect and/or analyze the information for the agency.***

The statistical aspects of the design have been reviewed thoroughly by staff at the National Center for Educational Evaluation and Regional Assistance (NCEE) in the Institute of Education Sciences (IES), by anonymous reviewers from IES' technical support contractor, and members of the Technical Work Group. Among the Technical Work Group members, Michael Kamil, Professor of Education at Stanford University, reviewed statistical aspects of the study design.

The following individuals have worked closely in developing the statistical aspects of the design.

Name	Organization and Title	Telephone
Stephen Bell	Abt Associates, Principal Associate	301-634-1721
Howard Rolston	Abt Associates, Principal Associate	301-634-1820
Larry Bernstein	Abt Associates, Senior Associate	617-349-2620

The SERVE Center at the University of North Carolina at Greensboro, Abt Associates Inc., and Empirical Education Inc. will be responsible for data collection and data analysis.

References

Campbell, J. M., Bell, S. K., & Keith, L. K. (2001). Concurrent validity of the Peabody Picture Vocabulary Test-Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment, 8*(1), 85-94.

Coyne, M. D., Simmons, D. C., Kame'enui, E. J., & Stoolmiller, M. (2004). Teaching vocabulary during shared storybook readings: An examination of differential effects. Except ionality, 12, 145-162.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4)*. Bloomington, MN: Pearson Assessments.

Graue, E. (1999). Diverse perspectives in kindergarten contexts and practices. In R. C. Pianta & M. J. Cox (Eds). The transition to kindergarten. Baltimore: Paul H. Brookes Publishing.

Guarino, C. M., Hamilton, L.S., Lockwood, J. R., & Rathburn, A. H. (2006). Teacher qualifications, instructional practices, and reading and mathematics gains of kindergarteners (NCES 2006-031). U.S. Department of Education. Washington, DC: National Center for Educational Statistics.

Heaviside, S., & Farris, E. (1993). *Public school kindergarten teachers' views on children's readiness for school* (NCES 93-410). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs* (3rd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Marvel, J., Lyter, D. M., Peltola, P., Strizek, G.A., and Morton, B.A. (2007). *Teacher Attrition and Mobility: Results from the 2004–05 Teacher Follow-up Survey* (NCES 2007–307). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S., & Ruston, H. (2006). Performance on the PPVT-III and the EVT: Applicability of the measures with African-American and European-American Preschool children. *Language, Hearing, and Speech Services in the Schools, 37*, 17-27.

Robbins , C., & Ehri , L.C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. Journal of Educational Psychology, 86(1), 54-64.

Smith, J., Brooks-Gunn, J., & Klebanov, P. (1997). Consequences of living in poverty for young children's cognitive and verbal ability and early school achievement. In G. Duncan & J. Books-Gunn (Eds.), *Consequences of growing up poor* (pp. 132-189). NY: Russell Sage Foundation.

Storch, S. A., Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. Developmental Psychology, 38(6), 934-947.

Tabors, P. O., Snow, C. E., & Dickinson, D. K. (2001). Homes and schools together: Supporting language and literacy development (pp. 313-334). In Dickinson, D.K., & Tabors, P.O. (Eds.)., *Beginning literacy with language: Young children learning at home and school.* Baltimore, MD: Paul H. Brookes Publishing Co.

Williams, K. T. (2007). *Expressive Vocabulary Test-Second Edition (EVT-2)*. Bloomington, MN: Pearson Assessments.

Woodcock, R. M. (1998). *Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU).* Bloomington, MN: Pearson Assessments.