NCEE 2010–4014

U.S. DEPARTMENT OF EDUCATION

The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB)

Kindergarten Final Evaluation Report

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

REL SOUTHEAST
Regional Educational Laboratory

The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB)

Kindergarten Final Evaluation Report

November 2010

Authors:

Barbara Goodson, Principal Investigator
Abt Associates Inc.*
Anne Wolf, Director of Evaluation
Abt Associates Inc.
Steve Bell, Task 2 Methodological Leader
Abt Associates Inc.
Herb Turner, Technical consultant
ANALYTICA
Pamela B. Finney, Regional Educational Laboratory Southeast Research Management Leader
SERVE Center at the University of North Carolina at Greensboro

*Joan McLaughlin, Original Principal Investigator, Abt Associates

Project Officer:

Sandra Garcia
Institute of Education Sciences

NCEE 2010–4014
U.S. Department of Education

U.S. Department of Education
Arne Duncan

Secretary

Institute of Education Sciences

John Q. Easton

Director

National Center for Education Evaluation and Regional Assistance

Rebecca A. Maynard

Commissioner

November 2010

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06C0-0028 with Regional Educational Laboratory Southeast administered by the SERVE Center at the University of North Carolina at Greensboro.

Disclosure of potential conflict of interest[1]

None of the authors or other staff involved in the study from Abt, ANALYTICA, Empirical Education, the Regional Educational Laboratory-Southeast, SERVE Center at the University of North Carolina at Greensboro, or the University of Georgia has financial interests that could be affected by the content of this report.[1] No one on the nine-member Technical Working Group, convened twice annually by the research team to provide advice and guidance, has financial interests that could be affected by the study findings.

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

CONTENTS

BOXES

FIGURES

MAP

TABLES

SUMMARY

Improving the ability of at-risk children to read and comprehend text has been a high priority in education policy over the last two decades. Low levels of reading achievement have been related to low academic performance, and one critical factor in reading achievement is adequate vocabulary knowledge. Children from disadvantaged backgrounds often lack general and academic vocabulary to enable them to acquire knowledge and comprehend text when they learn to read.

State education departments, in discussions with Regional Educational Laboratory (REL) Southeast, identified low reading achievement as a critical issue for their students and expressed an interest in identifying effective strategies to promote the foundational skills in young students that might improve reading achievement. The Mississippi State Department of Education has focused specifically on interventions that might enhance students' vocabulary knowledge. The Mississippi state legislature established a high priority on meeting the early education needs of students in or near the Delta, a primarily rural area of the state with a high level of poverty and historically low performance on reading achievement. To address these concerns, the current study tests the impact of a kindergarten vocabulary instruction program on students' expressive vocabulary when used across a range of districts and schools in the Mississippi Delta by kindergarten teachers as a supplement to their regular instructional program. Previous research on the program showed that a preschool version of the curriculum was associated with improved student vocabulary acquisition but did not provide a test using methodologies that can establish causal relationships.

Kindergarten PAVEd for Success (K-PAVE) was selected to be tested in Mississippi for three reasons. First, there were only a small number of vocabulary interventions appropriate for this age group to be considered. Second, among these, PAVE—the preschool version of the intervention—was the only one for which an impact study had been completed that provided some evidence of effects. Third, K-PAVE was the only curriculum that had developed teacher training materials and a training protocol, which meant that it could be implemented with sufficient fidelity across a variety of districts and school settings. The experimental design of this evaluation addresses limitations of earlier research and ensures a valid basis for estimating the effect of K-PAVE, implemented across a range of settings in the real world, on vocabulary knowledge of students in kindergarten.

K-PAVE is built around three components that support the acquisition of vocabulary in young students: instruction on a large set of thematically related target words through provision of definitions, examples, and visual images and through embedded instruction using storybook reading, extension activities, and teacher conversation; Interactive Book Reading to build vocabulary and comprehension skills; and Adult-Child Conversations to build vocabulary and oral language skills.

K-PAVE was implemented as a supplement to the regular classroom literacy instruction. Teachers were given broad latitude to choose how to integrate K-PAVE into their classroom instruction, including conducting K-PAVE activities in multiple curriculum areas across the classroom day and week. Fidelity of K-PAVE implementation was evaluated using a rating system provided by the program developer and administered based on classroom observation. Results showed that there was substantial variation in fidelity of implementation across

classrooms, which is typical of interventions implemented across a range of settings in the real world. At the same time, most classrooms were observed to be implementing K-PAVE with sufficient fidelity to support impacts on students.

The primary research question for the study addressed the impact of K-PAVE on kindergarten students' expressive vocabulary. Secondary research questions addressed the impacts on kindergarten students' academic knowledge and listening comprehension. Although the study was concerned primarily with the impacts of K-PAVE on students, impacts on intermediate classroom instruction outcomes were also assessed to provide context for understanding potential impacts on students. The study addressed research questions about impacts on classroom instruction in vocabulary and comprehension support, instructional support, and emotional support. Finally, the study examined whether the introduction of K-PAVE had the unintended consequence of reducing the time spent on areas of literacy instruction other than vocabulary and comprehension (such as phonological awareness, alphabet knowledge, print concepts, and decoding).

The experimental design of this evaluation addresses limitations of earlier research and provides a valid basis for estimating the effect of K-PAVE on the vocabulary knowledge of kindergarten students. The study employs a cluster random assignment design that randomly assigned schools to an intervention or control condition. Two kindergarten classrooms in each school were randomly selected to be in the study. The final sample included 65 schools, 130 kindergarten teachers, and approximately 1,300 kindergarten students distributed across 35 districts in the Mississippi Delta region and surrounding area. All districts, schools, teachers, and students volunteered to participate in the study and were not randomly sampled from the universe of eligible schools in the region.

The study sample did not differ significantly from all eligible schools in the Mississippi Delta and surrounding areas on a set of measured characteristics (including region, percentage of students eligible for free and reduced-price meals, and school accountability measures relating to school performance classification and meeting annual expectations for growth in achievement). Study schools did differ from eligible schools on two characteristics: study schools had a greater percentage of African American students and were more likely to be located in small towns and less likely to be located in rural areas.

To be eligible for the study, schools had to have at least two kindergarten classes, at least two consenting kindergarten teachers willing to be selected for data collection, and at least 40% of students eligible for free and reduced-price meals. The random assignment produced two groups of schools that did not differ significantly on pre-intervention measures of socioeconomic mix, state classifications of school-level student achievement, reading initiatives, racial composition of student body, region, and locale. In addition, students in the two groups of schools did not differ significantly on pre-intervention measures of expressive vocabulary, academic knowledge, listening comprehension, and other characteristics, including socioeconomic status, race, age, special education status, and gender.

After receiving training on the K-PAVE intervention, kindergarten teachers in treatment schools implemented it in the 2008/09 school year. The K-PAVE program was designed as a 24-week supplement to the core language arts program used in each school. Kindergarten teachers in control schools also implemented their core language arts curriculum and received their district's regular professional development during 2008/09. Teachers in the intervention

condition received two days of initial group training in fall 2008 (one month into the school year), three follow-up telephone conference calls to discuss implementation issues and reinforce key aspects of the K-PAVE program, and two rounds of classroom observation and feedback on how to improve their implementation of K-PAVE—one immediately after the initial training and one two-thirds of the way through the 24-week intervention.

The impact of K-PAVE at the end of kindergarten was assessed for one primary student outcome—expressive vocabulary. The estimated impact of K-PAVE on expressive vocabulary was 1.60 points on a scale with a mean of 100 and a standard deviation of 15,[2] and the impact was statistically significant. The standardized effect size for this impact is 0.14. Translating this effect size into a difference in age-equivalent scores, students who received the K-PAVE intervention were one month ahead in vocabulary development at the end of kindergarten compared with students in the control group (See Chapter 4, p. 85, and Appendix W for a discussion of how impacts on students can be translated into differences in age-equivalent scores).

The impact of K-PAVE at the end of kindergarten was also assessed for two secondary student outcomes—academic knowledge and listening comprehension. The impact of K-PAVE on academic knowledge was found to be statistically significant, with a magnitude of 1.95 points[3] on an item response theory-based scale with a sample mean of 455 points and standard deviation of 13.5 points in the control group. The standardized effect size for this impact is 0.14. Translating this effect size into a difference in age-equivalent scores, students who received the K-PAVE intervention were one month ahead in academic knowledge at the end of kindergarten compared with students in the control group. K-PAVE did not cause a statistically detectable impact on kindergarten listening comprehension.

The impact of K-PAVE at the end of kindergarten was also assessed for three classroom instruction outcomes hypothesized to foster students' vocabulary development: vocabulary and comprehension support, instructional support, and emotional support. K-PAVE caused a positive and statistically significant impact on vocabulary and comprehension support, which includes the introduction of vocabulary words throughout the school day and the use of comprehension supports and open-ended questions during book reading. The magnitude of this impact, 0.82 standard deviations, is equivalent to providing comprehension support during book reading 12 more times, asking three more higher order questions during book reading, introducing three more words during book reading, and introducing one more word during other times of the day. K-PAVE did not cause a statistically detectable impact on instructional support or emotional support in the classroom.

K-PAVE did not cause a statistically detectable impact on the amount of instructional time spent on literacy in areas other than vocabulary and comprehension. Thus, K-PAVE did not cause teachers to provide more vocabulary and comprehension support at the expense of time spent on other areas of literacy instruction.

[2] The 95% confidence interval around the impact estimate is 0.4–2.8 points.
[3] The 95% confidence interval around the impact estimate is 0.2–3.7 points.

CHAPTER 1: INTRODUCTION AND STUDY OVERVIEW

ROLE OF VOCABULARY KNOWLEDGE IN READING COMPREHENSION

The importance of vocabulary in reading achievement has long been recognized. The National Reading Panel (2000) has identified vocabulary as one of five key aspects of literacy involved in the reading comprehension of skilled readers. Oral vocabulary occupies an important middle ground in learning to read, as students move from oral to written forms of words (National Reading Panel 2000). To understand the meaning of the text, the beginning reader must be able to map an oral representation of a known word to the written word (or what is sometimes called reading vocabulary). Thus, learners who have a larger set of oral vocabulary are more likely to be able to apply that knowledge to print material. Evidence from studies of students with reading problems indicates that language and vocabulary deficits are critical factors underlying reading problems (Catts, Hogan, & Adolf 2005; Storch & Whitehurst 2002; Vellutino, Tunmer, Jaccard, & Chen 2007). Many children who successfully learn to read in grade 1 or grade 2 are unable to understand the books they need to read by grade 3 or grade 4 because they lack adequate vocabulary (Chall, Jacobs, & Baldwin 1990; Chall & Conard 1991; Spira, Bracken, & Fischel 2005; Storch & Whitehurst 2002). Conversely, high-knowledge grade 3 students have been reported to have vocabularies about equal to the lowest performing grade 12 students, and high school seniors near the top of their class are reported to know about four times as many words as their lower performing classmates (Beck, McKeown, & Kucan 2002). These data are the underlying rationale for the two decades of research investigating the effectiveness of different instructional interventions to enhance young children's vocabulary knowledge.

As stated by researchers in the field:

> It is essential that teachers engage struggling readers in activities that foster vocabulary development. Although wide reading should be encouraged and facilitated, struggling readers need more than just time to read. They seem to have difficulty gleaning the meanings of words from context and benefit from having new words and concepts that are critical to their learning taught directly to them. (Strickland, Ganske, & Monroe 2002)

SELECTING A VOCABULARY INSTRUCTION PROGRAM

The current study provides a rigorous test of the effectiveness of the vocabulary instruction program Kindergarten PAVEd for Success (K-PAVE) (Hamilton & Schwanenflugel, under contract) at increasing the vocabulary knowledge of low-income kindergarten students. Identifying effective strategies that schools can use to promote vocabulary acquisition in young, at-risk students is a critical challenge, based on the research showing that many low-income children enter school with limited vocabulary, which has consequences for their subsequent literacy development, especially reading comprehension (Biemiller & Slonim 2001; Coyne, Simmons, Kame'enui, & Stoolmiller 2004).

K-PAVE is a recently developed program to promote students' knowledge of vocabulary through multiple pathways, including explicit and embedded instruction of a set of target

vocabulary words and incidental exposure to other, novel vocabulary words. The program is designed to train teachers to use enhanced vocabulary instructional practices regularly and systematically. It is a modification of the original preschool PAVE program (Schwanenflugel et al., in press), which was designed to enhance early literacy skills in preschool children.

K-PAVE has three key components, each with a set of recommended teaching strategies. The first component, Explicit Vocabulary Instruction (labeled "New Vehicles"), involves explicit instruction of the target vocabulary words using word-learning strategies, exposure to the vocabulary words embedded in storybooks through repeated reading, and hands-on activities to extend student understanding of the meaning of the target vocabulary words. The second component of the K-PAVE program, Interactive Book Reading (labeled "CAR Talk"), involves teacher engagement of children during story reading through questions that promote comprehension and oral language skills. The third component, Adult-Child Conversations (labeled "Building Bridges"), involves frequent teacher conversations with individual or small groups of students to provide an opportunity for the teacher to use new vocabulary and for the students to increase their productive use of new vocabulary and their oral language skills generally.

The preschool program, PAVE, includes strategies not only to support vocabulary learning but also to enhance print knowledge and phonological awareness. K-PAVE does not include all of the components of PAVE that were implemented in the previous quasi-experimental study of preschool classrooms in Georgia public schools (Schwanenflugel et al. in press). The components that focused on the alphabet, phonological awareness, and uses of print were not included as part of K-PAVE. Schwanenflugel and her colleagues at the University of Georgia noted that, unlike in preschool, nearly all kindergarten teachers emphasize these components as part of general literacy instruction.[4] Other alterations to make PAVE appropriate for kindergarten included adaptations to the target vocabulary words and associated storybooks and to teacher training and support activities, to make it practical to take K-PAVE to scale. These changes are described in detail in Chapter 3.

K-PAVE was selected for testing for three primary reasons. First, there were only a small number of vocabulary interventions appropriate for this age group to be considered. Second, among these, PAVE—the preschool version of the intervention—was the only one for which an impact study had been completed that provided some evidence of effects (see below). Third, K-PAVE was the only curriculum that had developed teacher training materials and a training protocol. Therefore, to address the expressed desire of the Mississippi State Department of Education for a kindergarten vocabulary intervention, K-PAVE was selected for testing even though it represents an untested modification of the tested curriculum (PAVE) and its evidence of effectiveness was based on a quasi-experimental evaluation with some design limitations (described below). The earlier study of PAVE lacked the appropriate control groups to generate strong evidence about the effects of the program. As a randomized trial, the current study will provide a rigorous test of the effectiveness of K-PAVE in schools serving primarily low-income students.

In the current study, K-PAVE was implemented as a supplement to the regular kindergarten curriculum. Each week of the 24-week curriculum is organized around a

[4] In fact, the phonological awareness program adopted by PAVE was a popular kindergarten program called Phonological Awareness for Young Children (Adams, Foorman, Lundberg, & Beeler 1998).

vocabulary unit consisting of 10 thematically linked target words. The target words were selected to align with the themes in the Mississippi state science and social studies frameworks for kindergarten. Teacher training included initial group training, three follow-up telephone conferences over the 24-week program, and classroom visits to observe teachers implementing the curriculum and to provide remediation for teachers who were not implementing the curriculum practices with fidelity to the model. In this study, the group training and follow-up telephone conferences were led by the curriculum developer and a team from the University of Georgia. The classroom observations and remediation were conducted by a team from Regional Educational Laboratory Southeast, overseen by the curriculum developer.

This study is the first randomized test of K-PAVE. The previous study of the original PAVE preschool program evaluated the impact of PAVE using a quasi-experimental design that compared the intervention group with a comparison group and reported statistically significant differences in vocabulary knowledge (Schwanenflugel et al., in press). The intervention group students were from 18 volunteer schools (31 classrooms) in two counties. The 18 schools were randomly assigned to one of four PAVE treatment conditions, two that included the three components of the current K-PAVE intervention and two that did not include the Explicit Vocabulary Instruction component. Student participants were 180 boys and 165 girls ($n = 350$[5]) attending a full-day program for 4-year-olds (mean age = 4 years, 6 months, standard deviation = 4 months). According to parental report, 56% were African American, 40% European-American, 2% multiracial, 1% Hispanic, 1% Asian-American, and 1% not reported. Sixty percent of the children received free or reduced-price meals based on the federal formula for eligibility. Only children who were native English speakers according to parental report were included in the study. The sample included approximately 9% of children with identified special needs. One school in a neighboring county identified as being demographically similar to the treatment counties was recruited as the comparison. No data were reported on the baseline equivalence of the schools in the four treatment conditions and the comparison school; however, differences in children's baseline performance level were controlled in models examining the effects of PAVE on student outcomes.[6] On average, children in treatment schools that received the vocabulary components of the PAVE intervention scored significantly higher on expressive vocabulary than did children in the comparison school ($p < .01$). The standardized effect size ranged from 0.29 to 0.41, depending on which other PAVE components were part of the condition.

The findings suggesting an impact of PAVE on the vocabulary knowledge of young students from low-income households provide some evidence that the model builds vocabulary in young students. These results have to be interpreted cautiously since the design of the impact study raises concerns of potential selection bias in the assignment of schools to condition, statistical conclusion validity because of the small sample sizes, and internal validity because only a single school served as the control group. Nevertheless, in light of the paucity of appropriate curricula for kindergarten vocabulary instruction, the current study selected K-PAVE as the best candidate for an effectiveness trial because of the positive effects found in the earlier

[5] Data on gender were missing for five students.

[6] Pretest scores were entered as a level 1 predictor to adjust for the within-classroom variance among children in their initial skills on each outcome measure. Analyses of the equivalence of the groups indicated that there were baseline differences in child ethnicity and economic status (as determined by free and reduced-price meal rates) across the test conditions. Level 1 variables included these demographic variables and pretest scores (grand mean centered). Level 1 variables were dropped from the model if they did not account for statistically significant classroom-level variation.

study and because of the stage of development of the curriculum and associated materials. The current study provides a stronger test of the curriculum approach when implemented in kindergarten classrooms.

The current study is an effectiveness trial of K-PAVE because the curriculum is being tested under typical, real-world conditions rather than under optimal conditions. The study tests the impact of K-PAVE when implemented by kindergarten teachers as a supplement to their regular kindergarten instructional environment. Thus, school districts that are interested in implementing the curriculum could replicate the conditions under which the impacts of K-PAVE were tested in the current study, including the instructional materials and professional development. Even though the targeted vocabulary and associated materials and books were selected to align with the Mississippi science and social studies frameworks, the vocabulary is appropriate for kindergarten children, is likely to be consistent with kindergarten standards in other districts, and could be easily adapted by other districts. K-PAVE is not currently available from a publisher, but the developer has negotiated a contract with a publisher to make it available. Those who are interested in K-PAVE should contact the developer for more information.

THEORY OF CHANGE FOR K-PAVE

Although improvement in students' vocabulary knowledge at the end of kindergarten is the primary outcome for the study, the intervention model extends beyond explicit vocabulary instruction. K-PAVE's focus on conversation and interactive book reading is hypothesized to affect students' academic knowledge and comprehension as well as their vocabulary knowledge. Learning vocabulary and acquiring general knowledge are related; each supports the other. For example, when explaining the meaning of words to children, adults often make connections to students' existing knowledge. At the same time, learning new information, perhaps as part of a school unit on new academic content, provides opportunities for learning new vocabulary. Furthermore, increasing students' vocabulary and knowledge about the world is a pathway to skills in comprehending spoken language. Also, as students learn to read, their vocabulary and background knowledge support their comprehension of print. For these reasons, K-PAVE also is hypothesized to have impacts on secondary student outcomes of academic knowledge and listening comprehension.

Further, K-PAVE is hypothesized to lead to sustained advantages in vocabulary knowledge through grade 1—increased vocabulary acquisition in kindergarten is assumed to have continuing positive effects on vocabulary acquisition in the next school year, even in the absence of the intervention in grade 1. The logic model for sustained effects of K-PAVE in grade 1 assumes that increased vocabulary knowledge at the end of kindergarten will lead to impacts at the end of grade 1 that go beyond vocabulary to comprehension and academic knowledge.

The pathway by which the K-PAVE intervention is hypothesized to enhance students' vocabulary knowledge is through impacts on kindergarten teachers' instructional practices. Teachers are trained to implement specific K-PAVE instructional strategies aimed at building vocabulary, which are hypothesized to result in stronger vocabulary knowledge for students. To test this theory of change, the study includes estimates of impacts on teachers' instructional practices.

EXPERIMENTAL AND QUASI-EXPERIMENTAL EVIDENCE OF IMPACTS OF VOCABULARY INSTRUCTION

K-PAVE builds on the body of literature about effective strategies for promoting vocabulary knowledge in young students. Although research on the impacts of Explicit Vocabulary Instruction has focused on students in grade 3 and higher (Baumann, Kame'enui, & Ash 2003), gaps in vocabulary evident at school entry (Biemiller & Slonim 2001) underscore the importance of vocabulary instruction before grade 3. The research suggests that vocabulary can be improved through systematic instructional strategies to promote vocabulary learning in students below grade 3. In general, vocabulary interventions in preschool and kindergarten embed vocabulary instruction in storybook reading, part of most programming for students in this age group and a primary source of new vocabulary for young children. Although many of the interventions embed vocabulary instruction in storybook reading, the studies test such variants as whether stories are read once or multiple times, whether meanings are provided directly for vocabulary words encountered in reading, and whether students have opportunities to interact with the book during reading.

The studies also share design characteristics. Typically, the studies used pre-post designs, which test students on their knowledge of instructed words that they were exposed to in the storybooks and present results as gains in student knowledge of word meanings. Few studies also assess the effect of vocabulary instruction on untaught vocabulary, for example, by using a standardized receptive or expressive vocabulary measure. Also, the sample sizes in the studies tend to be small. These design features have limited the strength of the evidence on the impacts of vocabulary instruction on these outcomes for young students.

The research on vocabulary interventions has been summarized in two meta-analyses—the National Reading Panel (2000), which reviewed studies published before 1998, and the National Early Literacy Panel (2009), which reviewed studies published through 2003—and by individual researchers (see Biemiller & Boote 2007). Findings have been reported for different approaches to vocabulary instruction in storybook reading. Studies testing the impact of repeated reading of a story without explicit explanations of word meanings reported gains of 5%–9% in instructed words learned at the end of the intervention (Elley 1989; Robbins & Ehri 1994; Hargrave & Senechal 2000; Penno, Wilkinson, & Moore 2002; Brabham & Lynch-Brown 2002; Biemiller & Boote 2006).

A second set of studies tested the effect of augmenting the reading aloud with explicit explanation of the meaning of the words during the story reading. In general, these studies indicate that this strategy increases the effectiveness of the vocabulary instruction. For example, a set of studies that tested the effectiveness of a single reading of a story with explicit explanations of word meaning reported an average gain of 12% in instructed words learned (Senechal 1997; Senechal & Cornell 1993). Studies that examined the impact of repeated readings of a story with explanations of word meaning reported an average gain of 17% in word meanings known (Robbins & Ehri 1994; Senechal 1997; Senechal, Thomas, & Monker 1995; Hargrave & Senechal 2000; Penno, Wilkinson, & Moore 2002; Brabham & Lynch-Brown 2002; Biemiller & Boote 2006). Some of these studies compared immediate posttests with posttests six weeks to three months after the end of the vocabulary instruction and reported word knowledge to be about the same or higher at delayed testing (Senechal & Cornell 1993; Senechal, Thomas, & Monker 1995). Finally, studies that tested the effectiveness of involving students in interactive

word discussions during repeated story readings report average gains of 12% in instructed words learned (Brabham & Lynch-Brown 2002; Hargrave & Senechal 2000).

It has been pointed out that repeated reading of the same story does not expose students to additional contexts for increasing their understanding of target words (Beck & McKeown 2007). More recent studies have included follow-up extension activities or vocabulary reviews as means to increase vocabulary knowledge. Wasik and Bond (2001) randomly assigned four preschool teachers to two vocabulary instruction conditions: the intervention teachers were trained to use Interactive Book Reading techniques combined with book reading extension activities, and the control teachers continued with the regular classroom programming. The four classes in the study included 127 four-year-olds. At the end of the 15-week intervention, students taught by the intervention teachers had significantly higher scores on a standardized vocabulary test ($p < .001$) and on tests of receptive and expressive vocabulary developed for the study ($p < .01$).

Coyne et al. (2004) conducted an experiment to test the impact of embedded instruction with 96 at-risk kindergarten students from seven schools. Students were randomly assigned to one of three groups: a storybook intervention with embedded instruction, an intervention that focused on increasing phonologic and alphabetic skills, or a control group that received a module on sounds and letters from a commercial reading program. Students in all groups received 30 minutes of small group intervention each day for seven months, for a total of 108 instructional periods. The storybook intervention consisted of lessons developed to accompany 40 storybooks. Three target vocabulary words were taught explicitly from each storybook, and two storybooks were read twice each week. In addition, each week children were given opportunities to retell the stories using selected illustrations as prompts, with encouragement from teachers to use target vocabulary. At posttest, students in the storybook group scored significantly higher than students in the code-based and control groups on an experimenter developed expressive measure of explicitly taught vocabulary. The effect size was 0.73 for the contrast of the storybook and the code-based interventions and 0.85 for the contrast of the storybook intervention and the control group.[7]

A recent group of studies tested an instructional approach different from the embedded instruction using storybook reading. Extended vocabulary instruction or rich instruction is hypothesized to help students develop greater depth of vocabulary knowledge through exposure to multiple examples of vocabulary words in multiple contexts. This, in turn, is expected to help students process words more deeply by identifying and explaining appropriate and inappropriate uses and generally by providing extended opportunities to discuss and interact with words in multiple contexts in addition to story reading.

Beck and McKeown (2007) conducted two quasi-experimental studies to test the effectiveness for students in kindergarten and grade 1 of a treatment called Text Talk, which provided opportunities for rich language development through discussion of complex narratives in storybooks selected to be conceptually challenging. In the first study, the intervention consisted of book read-alouds and associated vocabulary instruction over a 10-week period. The

[7] The effect sizes reported in many of the studies on vocabulary instruction are large, some more than one standard deviation. The large effect sizes may be related to the fact that the outcomes are based on gain scores or on experimenter-developed measures, often without a comparison group, and to the fact that research has shown that quasi-experiments, on average, produce more optimistic estimates of effect sizes than do experiments.

study was conducted in eight classrooms in an urban school district with a low-income, predominantly African American population. Two classrooms from each grade were designated as intervention (implementing Text Talk) or as control (implementing daily read-alouds as part of the regular school reading curriculum). The student sample included 98 students in the eight classrooms. In both grades the intervention students learned significantly more of the 22 instructed words. The mean gain for the students in the intervention classes was 5.6 words, compared with 1.0 word for students in the control group classes ($p < .000$; effect size = 1.17). Comparable gains in grade 1 were 3.6 words for students in intervention classes and 1.7 words for students in control classes ($p < .01$; effect size = 0.74).

The second study tested whether students learned more vocabulary if teachers spent more time on the intervention. The sample consisted of three kindergarten classrooms (36 students) and three grade 1 classrooms (40 students), half designated to implement the same intervention as in the first study and half designated to implement the intervention but with additional instructional time. (Observations showed that Text Talk alone resulted in five encounters per vocabulary word, while the enhanced version of the intervention resulted in 20 encounters.) In both grades, students in the enhanced instruction classrooms made significantly greater gains in the number of instructed words learned ($p < .001$). Average gains in kindergarten were 8.2 of the 42 instructed words for students in the enhanced intervention classrooms and 2.5 words for students in the regular intervention classrooms; comparable gains in grade 1 were 6.9 words and 3.8 words.

Coyne and colleagues have conducted a set of experimental studies comparing different methods of vocabulary instruction with kindergarten students. Both the embedded and extended instruction conditions involved direct teaching of the meaning of target vocabulary words in the context of story reading, but the extended instruction also included opportunities for students to interact with and discuss target words in varied contexts outside of the story as a way to extend their understanding of the words. Students were taught three vocabulary words using each of these methods—embedded, extended, and incidental exposure. At the end of the intervention students were tested on their receptive knowledge, expressive knowledge, and depth of knowledge of the target words in multiple contexts. In one study, 32 kindergarten students received extended and embedded instruction. Students scored significantly higher at immediate posttest on all three measures—expressive and receptive definitions and context—on words that received extended instruction compared with words that received incidental exposure ($p < .001$ for the three outcomes, with effect sizes of 2.27 for expressive definitions, 1.00 for receptive definitions, and 1.02 for context) (Coyne, McCoach, and Kapp 2007). In a second study with another sample of 32 kindergarten students that compared extended and embedded instruction, students scored significantly higher on all three outcome measures on words that received extended instruction compared with words that received embedded instruction ($p < .001$ on all three outcomes, effect sizes of 1.70, 0.99, and 1.12, respectively) (Coyne, McCoach, and Kapp 2007).

A third study that compared all three instructional conditions with the same sample of 42 kindergarten students reported findings that were consistent with the earlier studies (Coyne, McCoach, Loftus, Zipoli, and Kapp 2009). On tests of receptive and expressive definitions, students had significantly higher mean scores for words taught using extended instruction than for words taught with embedded instruction ($p < .01$; effect size of 0.70 for receptive definitions and 1.34 for expressive definitions). Mean scores were significantly higher on vocabulary tests

of words taught using either of the intervention approaches than for words taught through incidental exposure. For extended instruction, the difference was significant at $p < .01$ for both receptive and expressive definitions, with effect sizes of 0.97 and 2.57 respectively. For embedded instruction, the difference was statistically significant at $p < .05$ for receptive definitions (effect size of .24) and at $p < .01$ for expressive definitions (effect size of 0.87). On the measure of word knowledge at the end of the intervention, mean scores were significantly higher for extended instruction than for embedded instruction, for both full knowledge ($p < .01$, effect size of 0.38) and for partial knowledge ($p < .01$, effect size of 0.56). Mean scores also were significantly higher for each of the intervention approaches compared with incidental exposure. For extended instruction, the difference was significant for measures of full and partial knowledge (effect sizes of 0.91 and 1.07). For embedded instruction, the difference also was significant for the measures of full and partial knowledge (effect sizes of 0.63 and 0.49).

Another recent study examined a language and literacy intervention that shares a number of instructional components with K-PAVE and reported impacts on children's vocabulary acquisition (Landry, Anthony, Swank, & Monseque-Bailey 2009). The study was primarily a test of four professional development programs, all of which focused on improving the language and literacy instruction among teachers of at-risk preschool children. The study employed a large sample of 262 classrooms in 158 schools across four states, to test the feasibility of taking intensive professional development to scale. Classrooms were randomly assigned to business as usual or to one of four professional development conditions, which varied in terms of receipt of in-class mentoring and detailed instructionally linked feedback concerning children's progress in language and literacy. On a measure of expressive vocabulary, the children whose teachers received the most intensive professional development (i.e., both mentoring and feedback) had statistically significantly higher scores at posttest (effect size = 0.19).

There is a growing research base on instructional strategies that have been found to promote vocabulary learning. The small set of experimental and quasi-experimental studies in this area, in addition to the pre-post studies, suggest that early elementary school students' vocabulary can be improved through interventions involving instructional strategies to promote vocabulary learning. Effective strategies include explicitly instructing students in vocabulary words, embedding vocabulary instruction in repeated reading of the same book and providing opportunities for students to actively participate in book reading, extending students' opportunities to engage with word meanings beyond book reading (explaining meanings, using words in new contexts, discriminating among examples of word meanings, allowing students to provide their own meanings) and combining strategies for embedded and extended strategies for teaching word meanings.

With the exception of the recent work by Landry et al. (2009), research on instructional strategies tends to involve small-scale efficacy studies, however, testing short-term interventions under ideal conditions, with developer support to ensure optimal implementation fidelity, and often using developer-created outcome measures that test students only on intervention target words. The current study of K-PAVE is the first large-scale effectiveness trial testing a vocabulary intervention in kindergarten under typical conditions over the course of a full school year, using a standardized vocabulary outcome measure.

STUDY TARGETS THE MISSISSIPPI DELTA REGION TO ADDRESS LOW STUDENT READING ACHIEVEMENT

Research supports that vocabulary acquisition is related to a child's early environment. Children from households that live in poverty are more likely to enter school with poorly developed language skills, including vocabulary, than are children from households with more resources (Smith, Brooks-Gunn, & Klebanov 1997). A large disparity in vocabulary is apparent at least as early as age 3. At this age children from middle-class households have vocabularies of approximately 1,000 words, while children from households living in poverty have vocabularies half that size (Hart & Risley 1995). The initial disparity continues through elementary school. By grade 1, children from middle-class households know approximately 5,000 words, while children from households living in poverty know only about 3,000. By grade 4, the respective vocabularies are 16,000 words and 11,000 words (White, Graves, & Slater 1990; Beck, McKeown, & Kucan 2002). These trends highlight the need for instructional interventions to accelerate vocabulary acquisition in young children (Biemiller 2001; Catts et al. 2005).

The Mississippi Delta region was selected as the target area for the study for three reasons: Students in the Delta are at increased risk for poor reading outcomes based on the high levels of family poverty in the region; students in the Delta have a history of low achievement scores, and the state legislature had established a high priority on meeting the early education needs of students in the Delta.[8]

The Delta area is primarily rural with a poverty rate more than twice the national rate (Bishaw & Iceland 2003).[9] According to the U.S. Bureau of the Census (2008), the poverty rate in 2007 for children under age 18 in the Delta region counties averaged 45.3%, with poverty rates in most Delta counties above 36% and as high as 57.6%. The overall poverty rate in 2007 for children under age 18 was 29.7% in Mississippi and 18.0% nationally. Mississippi's poverty rate is the highest of any state in the U.S.

For the Southeast Region as a whole, data from several psychometric studies find that in this region, children from households living in poverty have particularly low vocabulary scores, at about one standard deviation below the national average (Campbell, Bell, & Keith 2001; Restrepo et al. 2006). Oral language deficits manifest themselves in academic difficulties as these children transition from learning to read to reading to learn. On average, 18% of grade 3 and 4 students in Alabama, Florida, Georgia, Mississippi, and South Carolina do not meet state reading standards. By middle school the proportion is 32%. The rate in middle school is more pronounced for African American students (41%) and economically disadvantaged students (40%). Mississippi ranks 50th among U.S. states on grade 4 reading scores and 49th on grade 8 reading scores on the National Assessment of Educational Progress (Quality Counts 2008). It is in this context that Mississippi passed the Delta Revitalization Act of 2006, which focuses on revitalizing the Delta region on several fronts, including an emphasis on meeting the early education needs of students through grade 3.

---

[8] Mississippi is also the focus of the study because it is one of the target states in the region served by the Regional Educational Laboratory-Southeast, 1 of 10 laboratories funded by the U.S. Department of Education's Institute of Education Sciences. Southeast Region states include Alabama, Florida, Georgia, Mississippi, North Carolina, and South Carolina.

[9] See Appendix A for map of Mississippi showing the Delta region.

STUDY DESIGN

Using an experimental design, the study addresses whether K-PAVE has impacts when implemented under real-world rather than optimal conditions. The study was conducted in a sample of kindergarten classrooms in multiple school districts in the Mississippi Delta and surrounding areas. Based on a statistical power analysis, the sample includes 33 school districts, 65 schools, 130 kindergarten classrooms (two per school), and 1,296 students (about 20 per school). The study used a cluster random assignment design that assigned schools either to the K-PAVE intervention as a supplement to the usual classroom instructional program or to the usual instructional program not supplemented by K-PAVE.

The study is also designed as an effectiveness study, to test whether the intervention can be successfully implemented and can improve student vocabulary acquisition under real-world conditions. Kindergarten teachers in the intervention group received the K-PAVE intervention training in fall of the 2008/09 school year. Kindergarten teachers in the control group received their district's usual professional development during the intervention year and were offered the K-PAVE training the following school year. Kindergarten students in both intervention and control schools were assessed in fall 2008, prior to intervention implementation, and in spring 2009, to examine impacts of K-PAVE on vocabulary development and related outcomes. Classroom instruction in both intervention and control schools was observed in fall 2008, prior to intervention implementation, and in spring 2009, to examine the impacts of K-PAVE on kindergarten instructional practices.

Study objectives

The primary objective of the K-PAVE intervention is to improve students' expressive vocabulary in kindergarten. Because the intervention seeks to enhance students' vocabulary not only through explicit instruction but also through teachers' informal conversations with students and interactive book reading, K-PAVE is also hypothesized to have a secondary impact on children's general knowledge and listening comprehension. To test these hypotheses, the study addresses one primary research question about the impact of K-PAVE on kindergarten students' expressive vocabulary and a secondary research question on the impacts on kindergarten students' academic knowledge and listening comprehension:

1. What is the impact of K-PAVE on students' expressive vocabulary in kindergarten?
2. What is the impact of K-PAVE on students' listening comprehension and academic knowledge in kindergarten?

The pathway by which K-PAVE is hypothesized to improve children's expressive vocabulary and, secondarily, their general knowledge and comprehension skills is through impacts on teachers' instructional practices. Impacts on intermediate classroom instruction outcomes are examined to provide context for understanding potential impacts on students. The study addresses two secondary research questions about impacts on classroom instruction. One question focuses on impacts on instruction in areas hypothesized to foster impacts on students—vocabulary and comprehension support, instructional support, and emotional support. The other question examines whether the introduction of K-PAVE has the unintended consequence of reducing the time spent on areas of literacy instruction other than vocabulary and comprehension (such as phonological awareness, alphabet knowledge, print concepts, and decoding).

3. What are the impacts of K-PAVE on kindergarten instructional practices, specifically vocabulary and comprehension support during book reading, instructional support, and emotional support?
4. Does implementation of explicit and embedded vocabulary teaching reduce classroom time for instruction in other areas of early literacy (such as concepts of print, phonological awareness, alphabetic knowledge, reading, writing, fluency, and spelling)? Specifically, do K-PAVE teachers spend less time on these nonvocabulary literacy teaching practices compared with control teachers?

The study also addresses a secondary research question about the hypothesized sustained effects of K-PAVE on vocabulary knowledge and other reading-related skills in grade 1. This component of the study will be conducted only if K-PAVE is found to have impacts on students' expressive vocabulary at the end of kindergarten. The associated research question asks:

5. What is the impact of K-PAVE on students' expressive vocabulary, listening comprehension, academic knowledge, and passage comprehension in grade 1?

Exploratory research questions to be addressed and reported separately

In a second report, the study will investigate several exploratory questions involving student outcomes that K-PAVE is also hypothesized to affect but for which there is no strong research basis for the hypotheses:

1. What is the impact of K-PAVE on students' lexical diversity in kindergarten?
2. Do intervention effects on children in kindergarten or in grade 1 vary with student characteristics (such as age, gender, and entry-level pretest standard score)?
3. What are the impacts of K-PAVE on kindergarten teachers' lexical diversity and instructional practices during book reading?

These analyses are intended to suggest directions for future research on vocabulary interventions. The analyses of impacts on the lexical diversity measures for both students and teachers are considered exploratory because the measure itself and the procedures for its measurement are relatively new.

The study will also explore differential impacts for subgroups of students, such as males and students with below average pretest scores. These analyses are considered exploratory not only because the study does not have a priori hypotheses about subgroup differences but also because the study is not powered to calculate reliable estimates of impacts on subgroups of students. Finally, exploratory analyses will be conducted of the impacts of K-PAVE on instructional practices embedded within the instructional practices examined in the confirmatory analyses. For example, we examine impacts on a broad measure of vocabulary and comprehension support in the classroom as part of the confirmatory impact analysis. Exploratory analyses will examine elements of broader vocabulary and comprehension support, such as asking higher-order questions during book reading.

Contribution of the current study to research on vocabulary instruction

The current study is the first randomized study of the K-PAVE vocabulary intervention and a rigorous test of the impacts of vocabulary instruction on young students who have not yet encountered formal reading instruction. The results provide policymakers with important evidence on whether K-PAVE can produce impacts when implemented in school districts under the same real-world conditions.

The study provides evidence on two topics on which there is little empirical evidence: whether vocabulary interventions have impacts not only on vocabulary knowledge but also on broader student skills, including academic knowledge and comprehension, which are considered foundational skills for later reading achievement, and whether vocabulary instruction in kindergarten has sustained effects beyond the intervention period, in grade 1.

REPORT OVERVIEW

The remainder of the report details study design and methodology, implementation of the intervention, and impact results. Chapter 2 describes the study design and methodology, including sample recruitment, random assignment, data collection, outcome measures, response rates, analytic sample sizes, and data analysis methods. Chapter 3 discusses implementation of the intervention, including delivery and receipt of the intervention, variation in implementation, and fidelity to the program model. Chapter 4 presents the results of the impact analysis, which include estimated intent to treat impact findings for primary and secondary student outcomes. Impacts on intermediate classroom instruction outcomes are also presented in Chapter 4. Chapter 4 includes only analyses specified prior to any examination of outcome data, which are intended to address a priori hypotheses about intervention impacts. Chapter 5 summarizes the major findings of the study and design features that limit the generalizability of the results. In addition, technical appendices are included, which provide more information about measures, statistical models, imputation of missing data, and sensitivity analyses.

CHAPTER 2: STUDY DESIGN AND METHODOLOGY

OVERVIEW

This study examines the impact of the Kindergarten PAVEd for Success (K-PAVE) intervention on students' vocabulary using a cluster random assignment design that randomly assigns schools to intervention or control conditions (see Box 2.1 for a design summary). The rigorous experimental design addresses limitations of earlier research noted in Chapter 1 and provides a valid basis for answering the study's key research questions.

Box 2.1 Summary of kindergarten evaluation design

Intervention. The K-PAVE vocabulary intervention was selected as the study intervention based on positive outcomes found in preschool and the identified need in the region for improving children's vocabulary outcomes. The K-PAVE intervention involved 24 weekly units that supplemented the core language arts program. The K-PAVE intervention was implemented October 2008–April 2009.

Participants. Study participants included 35 districts, 65 schools, 130 kindergarten teachers, and approximately 1,300 kindergarten students. Schools were recruited from districts in the Mississippi Delta region and surrounding area that were willing to allow schools to participate and be randomly assigned to intervention or control conditions. Schools were recruited from among those with at least two kindergarten classes, at least two consenting kindergarten teachers willing to be selected for data collection, and at least 40% of students eligible for free or reduced-price meals. Students were eligible to participate if they were enrolled in the kindergarten classes within the first month of the 2008/09 school year, had parental permission to be randomly selected for data collection, were not learning English as a second language, and did not have a language or hearing impairment that would prevent them from being assessed.

Research design. Schools were randomly assigned for the 2008/09 school year to the intervention group that would use K-PAVE or to a control group that would not have access to K-PAVE until the following school year. All kindergarten teachers in each school were assigned to the same condition. Since K-PAVE was a supplemental program, all intervention and control teachers were using some other language arts program. The study used multistage random sampling, which randomly selects two kindergarten classrooms from each school from a pool of consenting teachers before random assignment. Subsequently, a random sample of 10 students from each randomly selected classroom was selected for data collection from the pool of students with parental permission. Students in the intervention and control schools were assessed before the start and after the completion of the 24-week intervention. In both intervention and control schools, classroom instruction was observed by raters who were part of the evaluation team and not involved in K-PAVE implementation; observations took place before the start and at the end of the 24-week intervention period. Additional data on intervention and control schools were collected from teacher questionnaires and district student records, and K-PAVE intervention staff from the SERVE Center at the University of North Carolina, Greensboro, provided data on participation in intervention training and observed classrooms to measure intervention implementation fidelity.

Outcomes. Impact estimates focused primarily on student expressive vocabulary, as measured by the Expressive Vocabulary Test–2nd edition (Williams 2007). Secondary student outcomes included academic knowledge, assessed using the Woodcock-Johnson III/Normative Update academic knowledge test (Woodcock, McGrew, Shrank & Mather 2007) and listening comprehension, assessed using the Kaufman Test of Educational Achievement listening comprehension test (Kaufman & Kaufman 2004). Intermediate classroom instruction outcomes included four measures: a measure of vocabulary and comprehension support created for this study from two classroom observation measures (Read Aloud Profile–Kindergarten; Vocabulary Record); instructional support (Classroom Assessment Scoring System, or CLASS; Pianta, LaParo, Hamre 2008); emotional support (CLASS); and the proportion of observed instructional time devoted to nonvocabulary-related language arts instruction (CLASS).

According to the cluster random assignment design, all kindergarten classes in each sample school that agreed to participate were assigned to the same condition. The intervention was implemented from October 2008–April 2009; kindergarten teachers in intervention schools

were trained on and implemented the K-PAVE intervention in the 2008/09 school year. The K-PAVE program was designed to supplement the core language arts program being used in each school. Kindergarten teachers in intervention schools implemented K-PAVE along with their core language arts curriculum. Kindergarten teachers in control schools implemented their core language arts curriculum and received their district's regular professional development during the intervention year.[10] Control teachers were offered the K-PAVE intervention training the following school year.[11] Because of the experimental design of the study, differences between intervention schools and control schools in student outcomes and in classroom instruction that exceed chance (that are statistically significant) can be attributed to the K-PAVE intervention.

Assigning all classes in a school to the same condition eliminates concerns about potential diffusion of effects across classrooms within a school that could occur when classrooms are randomly assigned within each school (intervention practices could diffuse from intervention teachers to other teachers in a school through discussion, observation, or other forms of cross-teacher communication). Furthermore, having the intervention in all kindergarten classrooms in an intervention school is hypothesized to help support implementation. In the intervention schools, both kindergarten teachers and assistant teachers received the K-PAVE intervention training in the first year.

The study tests whether K-PAVE is effective when districts volunteer to participate and schools and teachers volunteer to implement the intervention. District superintendents, school principals, and teachers were all informed about the random assignment process and agreed to participate before schools were randomly assigned to the intervention or control condition. Consequently, the decision to participate was not influenced by whether a school received the intervention in 2008/09 or a year later. Eligible districts were under no obligation to participate; 86% (38 of 44) of districts recruited agreed to do so.[12] Once district superintendents agreed to participate, school principals were invited to participate; 85% of eligible schools in participating districts agreed to participate. Individual teachers could also decline to participate in the study. On average, 91% of teachers in each study school agreed to participate. The voluntary nature of the study and the fact that teachers, schools, and districts were participating by choice could mean that the impacts might differ from those that would result if a district mandated K-PAVE.

SAMPLE RECRUITMENT AND RANDOM ASSIGNMENT

The study took place in a rural area of the Mississippi Delta and in surrounding areas sharing characteristics with the Delta (see map in Appendix A), including high rates of poverty, low student achievement, and predominantly rural and African American communities. The target population for the study was all elementary schools in the Mississippi Delta with at least two kindergarten classes. The Delta Revitalization Act of 2006 identified the counties that comprise the Mississippi Delta region.

---

[10] Control group teachers received no monetary compensation or instructional materials during the intervention year.

[11] K-PAVE is a 24-week intervention for kindergarten classrooms. The fact that control teachers may implement K-PAVE in their kindergarten classrooms during the year following the intervention year is not expected to contaminate grade 1 impacts. Students in control schools will be in grade 1, and the grade 1 teachers will not receive the K-PAVE training. Although teachers in the same grade often plan instruction together, teachers for different grades rarely do. Grade 1 teachers are not likely to implement the K-PAVE program in grade 1 classrooms.

[12] See the section of this chapter on sample recruitment for more details on the eligibility criteria for schools and districts.

Recruitment of eligible districts, schools, and teachers

Based on the statistical power analysis, the recruitment target was 60–70 schools, with two classrooms per school and 10 students per classroom. (See Appendix B for details on the statistical power analysis.) All school districts in the Delta were recruited to participate in the study.

To be eligible for recruitment, districts were originally required to meet the following criteria (based on information from school year 2007/08):

1. The district is part of the geographic area known as the Mississippi Delta, a region of high-poverty and low-achieving schools, with some pockets of relative affluence. (The median percentage of students eligible for free or reduced-price meals is 91%; however, in some schools, the percentage is as low as 40%.)
2. The district has at least one school with two kindergarten classes and is in a high-poverty community (with at least 40% of students receiving free or reduced-price lunch).
3. The district has at least one school that meets these criteria and is willing and able to allow two consenting kindergarten teachers to be randomly selected for observation and 20 students to be randomly selected for testing from among students with parental permission.
4. The district is willing to allow schools that are eligible and willing to participate in the study to be randomly assigned to intervention or control conditions.

In the Delta region at the time of recruitment, there were 32 school districts in high-poverty communities with 74 schools that had at least two kindergarten classrooms. To ensure that a sample of at least 60 schools could be obtained (the sample size goal based on power calculations), the sampling frame was expanded to include schools from districts contiguous to and sharing demographic characteristics with the Mississippi Delta region.[13] After including schools from districts neighboring the Delta, the sampling universe comprised 44 school districts with 94 schools that met the eligibility criteria. All kindergarten classes in targeted schools were full-day programs.

The characteristics of the final sample of schools in the Delta and those in the surrounding area did not differ significantly on any measured characteristics except that schools in the Delta had a higher percentage of African American students (median of 97%), on average, than did schools in the surrounding region (median of 84% outside; $t = -2.28$; $p=.03$).

Figure 2.1 presents the sample recruitment process and timeline. Recruitment of districts, schools, and teachers took place over January–August 2008. Of the 44 districts in the sampling

[13] The counties bordering the Delta region include 24 districts with 86 elementary schools. After a set of exclusions because of nonsimilarity of schools with the Delta sample, 12 districts and 20 schools remained for recruitment to the study. Exclusions were made for three reasons. First, 18 schools were excluded because they served fewer low-income children than the Delta schools (fewer than 40% of students eligible for free or reduced-price meals). Second, the 35 schools in the Jackson Public School District were excluded because Jackson, as a mid-size city, is a more urban setting than anywhere in the Delta. Third, the Mississippi Department of Education evaluated the match between schools in contiguous counties and those in the Delta. The department started with a list of districts and elementary schools in contiguous counties with at least 40% of students eligible for free or reduced-price meals, and they indicated, based on their insider's perspective, whether each school was similar to those in the Delta in demographic, poverty, and achievement data. Based on their ratings, 13 more schools were eliminated from the potential sampling frame.

universe, 38 superintendents (86%) agreed to participate in the study. Once a district superintendent agreed to participate, principals of all eligible schools in the district were recruited for the study. Once schools agreed to participate, all kindergarten teachers in the sample of schools were recruited. Teachers were recruited to participate before random assignment to ensure that their decision was not influenced by the school's random assignment status.

School recruitment took place in two phases, one phase before random assignment and one phase after random assignment but before schools were notified of their random assignment status. Schools were initially recruited from February 25, 2008 to June 2008. Random assignment of schools occurred at the end of July 2008. All eligible schools that agreed to participate and had begun the consent process were randomized. Schools that were in the randomization pool were contacted in July to confirm willingness to participate; obtain consent for all kindergarten teachers, including any teachers hired since the first phase of recruitment; and obtain any outstanding written consent forms not yet submitted by the school.

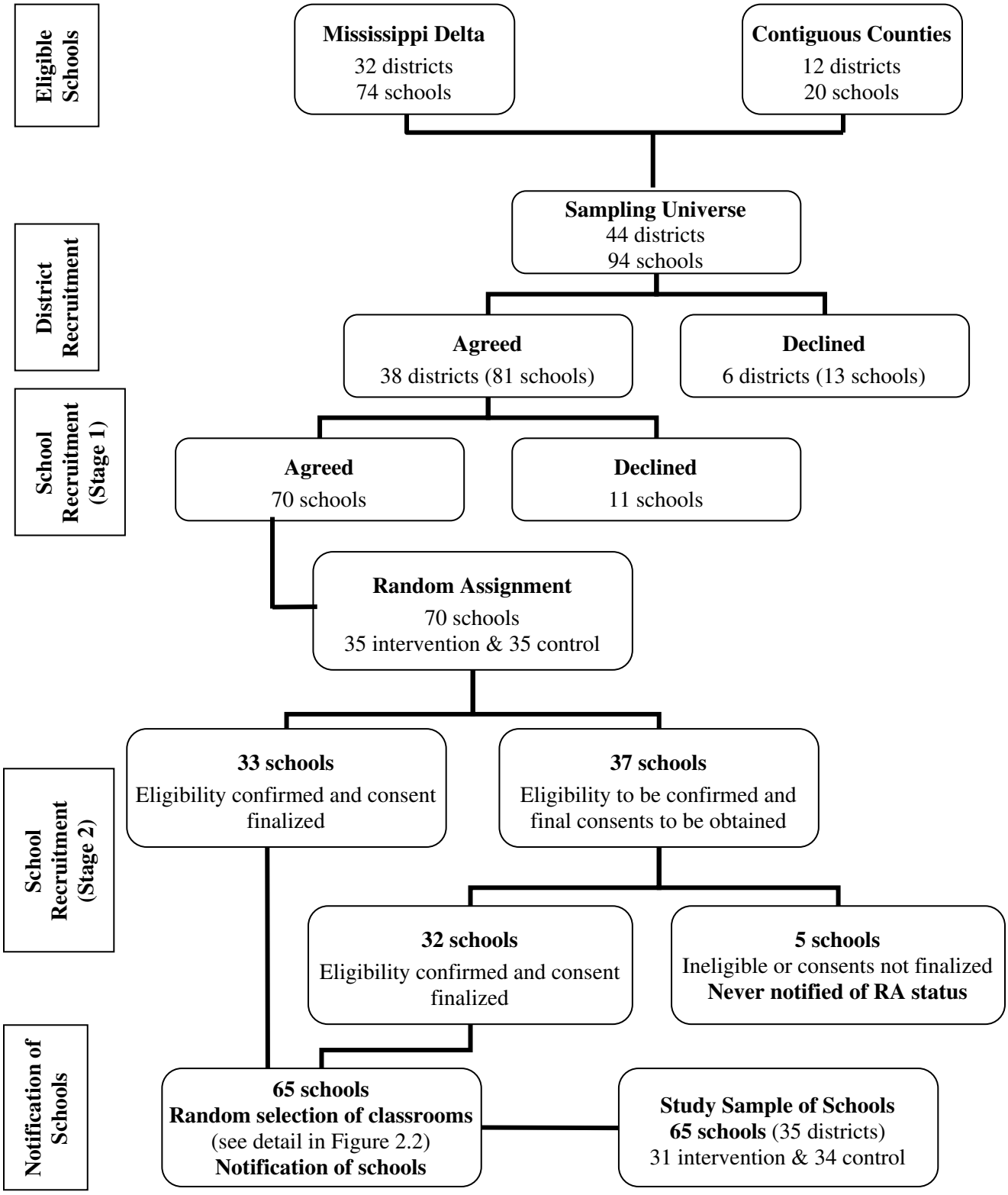
Of the 81 eligible in participating districts, 70 schools had indicated a willingness to participate in the study prior to randomization. All 70 schools were included in the pool of schools to be randomized, regardless of whether they had submitted all the written consent forms. Schools were randomized on July 24, 2008; 35 were assigned to the intervention group and 35 to the control group.

At the time of randomization, 33 schools had submitted all written consent forms and confirmed all staffing. For these schools, two kindergarten teachers from each school were randomly selected for data collection and, on July 25, 2008, letters were sent notifying these 33 schools of their random assignment status and the names of the teachers selected for data collection.

For the remaining 37 schools, efforts to confirm and complete the school and teacher consent process continued through August 7, 2008. *None of the 37 schools was notified of its random assignment status before all written consent forms were obtained.* By August 7, 2008:

- For 32 of the remaining 37 schools, all staffing was confirmed, and all the necessary written consent forms were submitted.
- The other five schools were excluded from the study because they did not meet eligibility criteria or because they were unwilling to participate in the study.

Figure 2.1. District and school recruitment process and timeline



Although all these schools were randomized, recruitment was not considered complete until schools confirmed their willingness and eligibility to participate and submitted all necessary written consent forms. All schools, including the five that were excluded based on unwillingness or ineligibility, went through the same confirmation process before being notified of their random assignment status. Because the schools were excluded before being notified of their assignment status, the integrity of the random assignment was maintained. That is, the school's decision not to participate was not influenced by its status as an intervention or control school, since the schools did not know to which group they were assigned.

The final sample included 65 schools: the 33 schools with complete written consent as of July, 24, 2008 and the 32 schools that had complete written consent as of August 7, 2008. The sample included 31 intervention schools and 34 control schools.[14]

Random assignment within blocks

Schools were placed into three blocks based on previous participation in reading initiatives.[15] Among the 65 schools in the study sample, 17 were Reading First schools, 5 had a Mississippi state reading initiative (Barksdale Reading or Mississippi Sufficiency), and 43 had a local district initiative or no reading initiative. Within the reading initiative blocks, schools were matched based on a set of school characteristics: School Performance Classification,[16] percentage of students receiving free or reduced-price meals, percentage of students who are African American, locale, and region (Delta or contiguous county). (Appendix C, which details the process of random assignment, includes a description of the matching within blocks.) Once

[14] Randomization did not result in equal numbers of schools in the intervention and control conditions because five schools that were randomized became nonparticipants (for ineligibility or incomplete consents) before notification of random assignment.

[15] Although within-district random assignment would have controlled for district characteristics, doing so would not have resulted in a sufficient sample size. Almost half the districts did not have enough schools to assign one to the intervention group and one to the control group. Of the 32 school districts in the Mississippi Delta region, 15 had only one elementary school with kindergarten. A within-district random assignment design would have reduced the sampling frame in the Delta from 74 schools to 59 schools. Furthermore, blocking based on substantive features of schools was preferred over blocking based on school districts, because experience with other reading initiatives is expected to be associated with differences in baseline classroom instructional practices. The strategy was thus intended to minimize differences between intervention and control schools in areas that are likely to be related to classroom instructional practices and student outcomes. Factors likely to drive impact levels can be quite heterogeneous within districts. The comparatively small number of schools and potentially large amount of variation within districts would tend to result in less comparable intervention and control groups than would random assignment with blocks based on prior experience with reading initiatives.

[16] The School Performance Classification is an annual classification based on students' performance on the state accountability test (Mississippi Curriculum Test [MCT]) administered to students in grades 3 and higher. Classifications include low performing, underperforming, successful, exemplary, and superior. Data on student achievement in the current year and on patterns of student growth from the prior year both contribute to the performance classification. A student's level of proficiency (i.e., minimal, basic, proficient, or advanced) in a given content area is determined based on his or her score on the MCT, with threshold scores dividing levels of proficiency. "Basic" proficiency is defined as "partial mastery of the content area knowledge and skills required for success at the next grade," and "proficient" is defined as "solid academic performance and mastery of the content area knowledge and skills required for success at the next grade" (http://www.mde.k12.ms.us/acad/OSA/gltp.html). A school's achievement level is classified based on the percentage of students scoring basic or higher and the percentage of students scoring proficient or higher. Data for students in all grades and all subject areas are combined. Schools are also rated on whether students meet or do not meet growth expectations on average.

matched based on these school characteristics, schools were randomly assigned to intervention or control conditions.

Random selection of classrooms

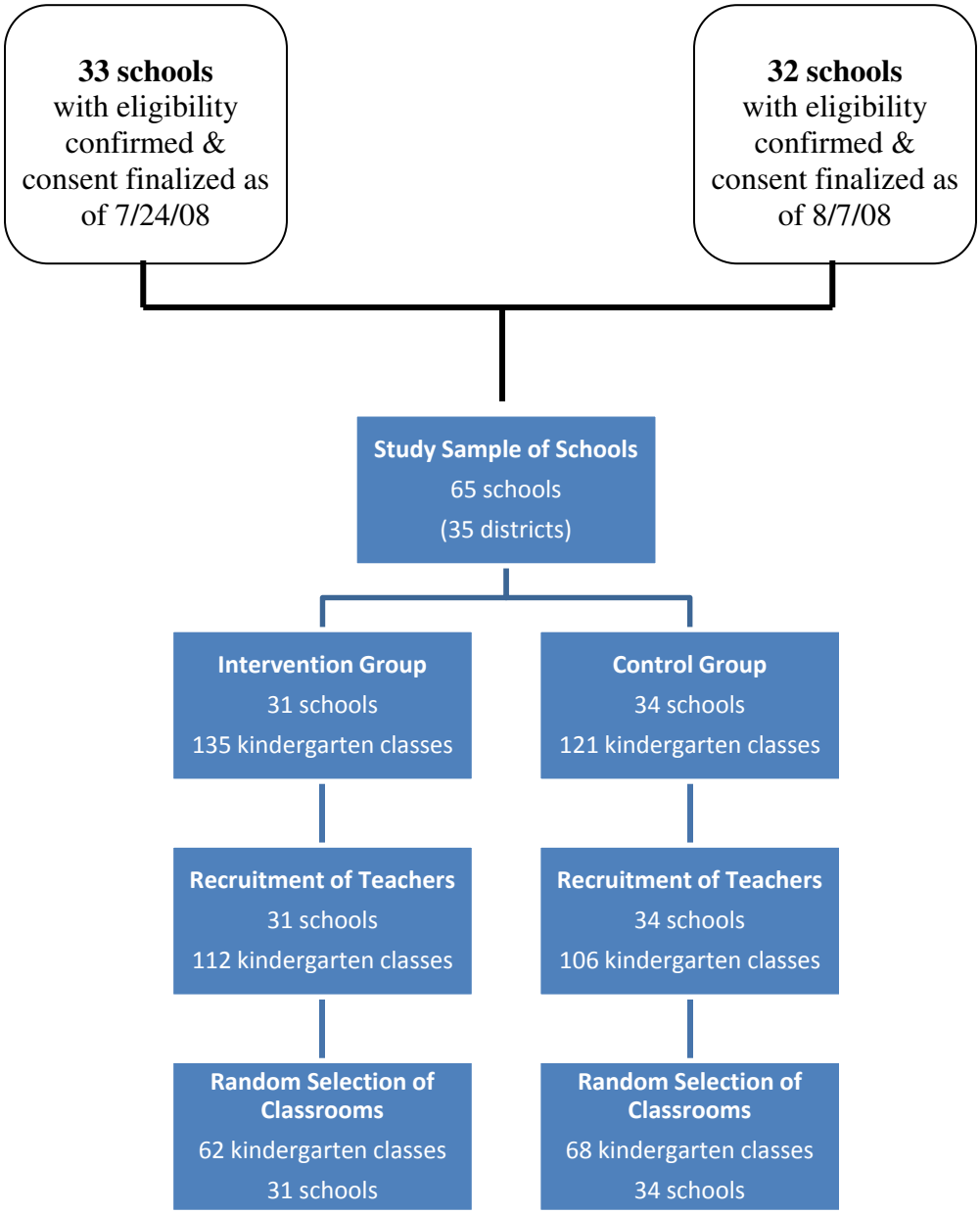
Figure 2.2 summarizes the schools that were recruited into the final study sample and randomized, recruitment of teachers, and random selection of classrooms. The 65 schools had a total of 256 full-day kindergarten classrooms. The average consent rate for teachers across the 65 schools was 91.5%.[17] Because at least two teachers from all 65 schools consented to participate in the study, no schools were lost due to lack of teacher consents. A total of 218 teachers consented to participate in the study.

From the pool of consenting teachers in each school, two were randomly selected to participate in data collection if there were more than two kindergarten classrooms in the school.[18] This random selection ensures that selected teachers are representative of teachers in their schools who are willing to participate in the study. For schools with only two kindergarten classrooms or with only two consenting kindergarten teachers (40% of the schools), both classrooms were selected. Schools were not notified of their assignment condition until after the random selection of teachers was completed.

[17] Teacher consent rates range from 20% to 100%. In 74% of schools, all kindergarten teachers consented to be in the study.

[18] There were more than two kindergarten classrooms in 66% of the schools—31% of schools had three kindergarten classrooms, 29% had 4–8 classrooms, and 6% had 11–15 classrooms. All kindergarten classes were full day.

Figure 2.2. Outcome of random assignment of schools and random selection of classrooms in the study sample



In the intervention schools all consenting teachers were offered the K-PAVE training, whether selected for data collection or not. All teachers who participated in the K-PAVE training received continuing education credits from Delta State University, based on the number of hours that teachers participated in the professional development activities.

Random selection of students

Based on statistical power calculations, the study set a goal of sampling 20 students from each participating school, equally distributed as 10 students per classroom. In the first month of the school year[19] the study sought permission from parents of all kindergarten students enrolled in the study classrooms to be assessed individually on their vocabulary and literacy skills in the fall and the spring. Theoretically, parents' decision to permit their child to be assessed could have been influenced by the school's intervention status, since recruitment of students occurred after the random assignment. However, implementation of the K-PAVE program had not yet begun at the time permission was sought from parents, and recruitment letters to parents did not provide information about schools' use of an intervention program. The study was described as attempting "to better understand how to help kindergarten children in Mississippi to develop the vocabulary skills they need to become successful readers." Therefore, it was considered unlikely that parents' willingness to permit their child to participate would have been affected by the school's assignment status.

If 20 or more students in a school had parental permission to participate, 20 students per school were randomly selected, preferably 10 students per classroom. If there were 20 or more students with parental permission but fewer than 10 students with parental permission in one of the study classrooms in a school, more students were randomly selected from the other study classroom in the school to achieve a total sample of 20 students in the school. If there were fewer than 20 total students with parental permission in the school, all students with permission to participate were selected with certainty. No adjustment for unequal numbers of students in classrooms or schools was made in the analysis.

The two study classrooms in a school were full-day classrooms, enrolling an average of 40 students, with enrollment ranging from 24 to 58 students. Parental permission was received for an average of 28 students per school (average permission rate of 73%), with the number of permissions ranging from 12 to 48 students per school. Only 11 schools (17%) received permission from fewer than 20 students. On average, 20 students per school were randomly selected to be assessed (average selection rate per school of 74%). From the 1,849 eligible students with parental permission, 1,276 students were randomly selected to be assessed—598 students in intervention schools and 678 students in control schools.

Appendix D illustrates the steps in recruiting and randomly selecting the student sample and shows the numbers of students involved at each step. (Also see Figure 2.3 in the section on response rates for data collection, which illustrates the flow of students through the study.) At the time of baseline data collection, 46 of the randomly selected students (23 intervention and 23 control) were not in school.[20] In their place 43 alternates (21 in intervention schools and 22 in control schools) were randomly selected for testing. The final analytic sample includes all of the

[19] Although school begins the first week of August in Mississippi, schools indicated that only about 60% of kindergarten students are enrolled by the start of school, with 20% of students arriving by the end of August and another 20% arriving after Labor Day. The study extended the period for obtaining parental permission forms to include late-arriving kindergarteners in the sample.

[20] Appendix E compares the characteristics of students who were not assessed at baseline with those who were. There were no statistically significant differences between the groups in terms of gender, eligibility for free or reduced-price meals, Individualized Education Programs status, and age. However, students who were not assessed at baseline were less likely to be African American than those who were tested ($t = 2.49$; $p = .007$). Of the 46 students not tested at baseline, 29 were tested at posttest.

students who were originally randomly selected as well as the 43 randomly selected alternates. In total, there were 1,319 students in the analytic sample—619 students in intervention schools and 700 students in control schools.

Characteristics of schools in the final study sample

The sample of schools that volunteered for the study was similar to the pool of eligible schools in the Delta and surrounding counties in terms of School Performance Classification (low performing, underperforming, successful, exemplary, and superior performing), expectations for annual growth in student achievement, and percentage of students eligible for free or reduced-price meals (Table 2.1). However, the schools that volunteered for the study differed from the pool of eligible schools on two characteristics. Sample schools had a greater percentage of African American students (mean of 85%) than all eligible schools (mean of 73%; $t = -2.17$; $p = .03$). In addition, a greater percentage of sample schools were located in small towns and a smaller percentage were located in rural areas than the pool of eligible schools ($x^2 = 6.53$; $p = .03$).

Table 2.1 Comparison of schools that agreed to participate and all eligible schools

	Agreed (study sample) 65 schools	Declined 28 schools	Sampling frame 93 schools	Test of difference[a]
Region				
Delta	83.1%	67.9%	78.5%	$x^2 = 2.69$
Outside	16.9%	32/1%	21.5%	$p = .10$
School Performance Classification (grades 3 and higher)				
Low or underperforming	20.7%	30.8%	23.8%	$x^2 = 4.30$
Successful	55.2%	30.8%	47.6%	$p = .12$
Exemplary or superior	24.1%	38.5%	28.6%	
Annual growth expectation (from the previous school year for grades 3 and higher)				
Met or exceeded	17.2%	11.5%	15.5%	$p = .75$
Not met	82.8%	88.5%	84.5%	
Eligibility for free or reduced-price meals				
96% - 100%	32.3%	21.4%	29.0%	
90% - 95%	29.2%	17.4%	25.8%	
70% - 89%	21.5%	28.6%	23.7%	
Less than 70%	16.7%	32.1%	21.5%	
Mean (standard deviation)	85.2% (16.3)	78.5% (19.0)	83.2% (17.3)	$t = -1.71$, $p = .09$
Percent of students who are African American				
96% - 100%	53.9%	32.1%	47.3%	
81% - 95%	21.5%	17.9%	20.4%	
Less than 81%	24.6%	50.0%	32.2%	
Mean (standard deviation)	85.0 (22.8)	73.1% (27.0)	81.4% (24.6)	$t = -2.17$, $p = .03$
Locale				
Rural	47.7%	67.9%	53.8%	$x^2 = 6.53$
Small town	36.9%	10.7%	29.0%	$p = .03$
Large town or fringe of mid-size city	15.4%	21.4%	17.2%	

Note: Distributions of school characteristics are assumed to be the same for cases with missing data as for cases with nonmissing data. Rates of missing data range from 0.0% to 10.8%.
a. Chi-square tests were used to test for differences between study schools and other eligible schools in School Performance Classification and school locale. *T*-tests were used to test for differences in eligibility for free or reduced-price meals and in the percentage of students who are African American; although categories are presented for these variables in the table, they are continuous variables. Fischer's exact test was used to test for differences in annual growth expectation because of small cell sizes.

Among the 65 schools in the sample, 26% were Reading First schools, 8% participated in a state reading initiative or pilot program (Mississippi Reading Sufficiency Program or Barksdale Reading Initiative), and 66% had either a local reading initiative or none (Table 2.2). There were 21% of schools that were classified as low performing or underperforming on the School Performance Classification for 2006/07, 55% that were classified as successful, and 24%that were classified as exemplary or superior. The majority of schools (83%) did not meet state expectations for annual growth in student achievement. The composition of the schools reflects the largely poor, African American population of the Delta and surrounding region. The median percentage of African American students in a school was 96% (not shown), and the median percentage of students who were eligible for free or reduced-price meals was 92% (not shown). Some 48% of the schools were in rural areas, 37% in small towns, and 15% in large towns or on the fringe of a city. The intervention and control schools were not significantly different at baseline on any of the characteristics in Table 2.2.

Table 2.2 Description of school sample

	Control 34 schools	Intervention 31 schools	Full sample 65 schools	Test of difference[a]
Reading initiatives				
Reading First/state reading initiative	35.3%	32.3%	33.9%	*p=.99*
Local or no initiatives	64.7%	67.7%	66.2%	
School Performance Classification (grades 3 and higher)				
Low or underperforming	19.4%	22.2%	20.7%	
Successful	58.1%	51.9%	55.2%	$p = .92$
Exemplary or superior	22.6%	25.9%	24.1%	
Annual growth expectation (from the previous school year for grades 3 and higher)				
Met or exceeded	16.1%	18.5%	17.2%	$p = .99$
Not met	83.9%	81.5%	82.8%	
Eligibility for free or reduced-price meals				
96% - 100%	26.5%	38.7%	32.3%	
90% - 95%	32.4%	25.8%	29.2%	
70% - 89%	23.5%	19.4	21.5%	
Less than 70%	17.7%	16.1%	16.7%	$t = 0.96$,
Mean (standard deviation)	84.8% (15.5)	85.6% (17.4)	85.2% (16.3)	$p = .34$
Percent of students who are African American				
96% - 100%	52.9%	54.8%	53.9%	
81% - 95%	20.6%	22.6%	21.5%	
Less than 81%	26.5%	22.6%	24.6%	$t = –0.34$,
Mean (standard deviation)	84.0% (15.2)	86.0% (20.3)	85.0 (22.8)	$p = .74$
Locale				
Rural	47.0%	48.4%	47.7%	
Small town	41.1%	32.3%	36.9%	$p = .62$
Large town or fringe of mid-size city	11.8%	19.4%	15.4%	
Region				
Delta	79.4%	87.1%	83.1%	

Outside	20.6%	12.9%	16.9%	p = .41

Note: Distributions of school characteristics are assumed to be the same for cases with missing data as for cases with nonmissing data. Rates of missing data range from 0.0% to 12.9%.
a. Fischer's exact test was used to test for intervention and control group differences in reading initiatives, School Performance Classification, and annual growth expectation because of small cell sizes. *T*-tests were used to test for intervention and control group differences in eligibility for free or reduced-price meals and in the percentage of students who are African American; although categories are presented in the table for these variables, they are continuous variables. Chi-square tests were used to test for differences in school locale.

As noted above, we blocked schools by reading initiative to ensure that intervention and control schools were balanced on this factor. We took this step because we expected that experience with Reading First or a state reading initiative would be related to classroom instructional practices and/or student outcomes. By ensuring that intervention and control schools were balanced with regard to reading initiatives, we were able to conclude that any impacts that we detect can be attributed to K-PAVE rather than to the Reading First program or a state reading program that may be present in certain schools. In addition, because experience with other reading initiatives may interact with how teachers implement K-PAVE and with its effectiveness for improving students' vocabulary outcomes, we plan to conduct a subgroup analysis in a future report to examine whether impacts of K-PAVE vary based on whether schools have the Reading First program or do not.

As noted above, K-PAVE was implemented as a supplement to literacy programs already in use in the intervention classrooms. According to reports from the schools, all study classrooms were using a commercial reading program. There was no statistically significant difference between intervention and control schools in the reading series used ($\chi^2 = 2.51$, p=.47; see Table 2.3). In both groups, more than 40% of schools reported using *Trophies*, 2005 Edition (Houghton Mifflin Harcourt School Publishers), in their kindergarten classrooms.

Table 2.3 Reading programs in place at baseline in intervention and control schools

Publisher/reading series	Intervention schools (percent)	Control schools (percent)	Test of difference
Trophies 2005 Edition (Houghton Mifflin Harcourt School Publishers)	43.3	41.2	$x^2 = 2.51$ $p = .47$
▪ *Treasures*, a Reading Language Arts Program, Grade K, Kindergarten System (MacMillan/McGraw-Hill, 2008) or ▪ *Houghton Mifflin Reading* (Houghton Mifflin Harcourt School Publishers, 2008)	26.7	23.5	
Other[a]	30.7	35.3	

Note: Sample includes all intervention schools ($n = 30$) and all control schools ($n = 34$).
a. Includes nine other curricula.
Source: Telephone survey of schools.

DATA COLLECTION

Overview of the units of data collection

To address the research questions, the study collected information on student performance, classroom instructional practices, student characteristics, teacher characteristics, school characteristics, and K-PAVE implementation. Table 2.4 provides a timeline of study activities, including data collection and intervention implementation. Data on student performance and classroom instructional practices were collected before the start of the K-PAVE intervention (for use as covariates in the impact analysis) and at the end of the school year. Student outcomes were measured at the end of the school year to estimate the impacts of K-PAVE on student expressive vocabulary and other vocabulary-related outcomes (academic knowledge and listening comprehension). Information about instructional practices at the end of the school year was used to examine how K-PAVE changed practices in literacy instruction and how teachers allocated their literacy instruction time (among vocabulary, comprehension, and other areas of literacy instruction). Data on student, teacher, and school characteristics were collected for use as covariates in the impact analysis. Information about the implementation of K-PAVE was used to assess the fidelity of implementation to program design.

Table 2.4 Timeline of study activities: data collection and intervention implementation

Date	**Student assessments**	**Observation of intervention and control classrooms**	**Student characteristics** (district administrative records)	**Survey of teacher characteristics**	**School characteristics** (online state database)	**K-PAVE fidelity rating** (intervention classrooms only)
Spring 2008					X	
August 2008		X		X		
September 2008	X	X		X		
October 2008	X[a]					
November 2008						
December 2008						
January 2009						
February 2009			X			X[b]
March 2009		X[c]	X			X[b]
April 2009	X[a]	X[c]	X			
May 2009	X		X			
June 2009			X			

Note: Shaded cells correspond to implementation of K-PAVE. The 24-week intervention period was staggered by one week for four groups of schools: October 6, 2008–April 10, 2009; October 13, 2008–April 17, 2009; October 20, 2008–April 24, 2009; October 27, 2008–May 1, 2009.

a. Baseline student assessments were completed before the start of the K-PAVE, and all posttest student assessments took place after completion of the K-PAVE intervention.

b. K-PAVE fidelity rating was conducted in intervention classrooms only, during weeks 17–20.

c. Posttest classroom observations were conducted in both intervention and control classrooms during weeks 22–24. All baseline classroom observations were completed before the start of the K-PAVE intervention.

Student assessment measures. Table 2.5 summarizes the student measures, the outcome variables derived from the measures, and the designation of each outcome as primary or secondary.[21] The primary research question examines the impact of K-PAVE on the outcome most directly targeted—students' expressive vocabulary in kindergarten. Other student outcomes in kindergarten and sustained student outcomes in grade 1 are considered secondary because they extend beyond the primary target of the intervention. Impacts on classroom instruction (discussed below) are considered secondary because they are the intermediate outcomes through which K-PAVE is hypothesized to affect students' expressive vocabulary.

Student outcomes were assessed for expressive vocabulary, academic knowledge, and listening comprehension. At baseline and again at the end of the kindergarten year, students were individually assessed for about 45 minutes by trained members of the evaluation team who were independent of the intervention and unaware of the school assignment to intervention or control conditions. Baseline assessments of all students in intervention schools and nearly all students in

[21] Outcomes are related to research questions that are driven by a priori hypotheses about the impacts of K-PAVE on student outcomes and classroom instruction. The research questions were specified prior to observation of the experimental outcomes and are not informed by any knowledge about the actual data.

control schools were completed before K-PAVE intervention training. There were 3.9% of students in control schools who were assessed one to three weeks later.[22] Posttest assessments for all students were conducted after completion of the 24-week intervention period.

Table 2.5 Child measures, outcome variables, and designations

Study area	Measure	Outcome variable	Status in analysis
Vocabulary knowledge	Expressive Vocabulary Test–2 (EVT–2)	Standard score	Primary
Academic knowledge	Woodcock-Johnson III/Normative Update (WJ–III/NU), Academic Knowledge subtest (science, humanities, social studies)	*W*-score, an item response theory–based scale score (all three content areas combined)	Secondary
Listening comprehension	Kaufman Test of Educational Achievement–II (KTEA–II), Listening Comprehension subtest	Standard score	Secondary

Expressive vocabulary. The primary measure of vocabulary acquisition in the study is the Expressive Vocabulary Test–2 (EVT–2; Williams 2007). Expressive vocabulary is the primary outcome in the impact analysis.

Assessing vocabulary development is a complex issue, because there are different kinds of vocabulary knowledge. Henriksen (1999) and Melka (1997) describe a receptive-expressive continuum where each word passes from receptive into productive use, with increasing exposure to the word (see Zareva, Schwanenflugel, & Nikolova 2005). In addition, Vermeer (2001) makes a distinction between breadth of vocabulary knowledge (number of words in the lexicon) and depth of such knowledge (how well those words are known). Expressive vocabulary captures both breadth and depth of vocabulary knowledge. Theoretically, receptive vocabulary represents the shallowest level of vocabulary knowledge. For example, children are likely to be able to select correctly from a multiple choice (such as the Peabody Picture Vocabulary Test, PPVT) with only partial word knowledge (Curtis 1987), whereas expressive vocabulary measures require a greater depth of knowledge not only to recognize but also to recall a word.

Although the PPVT, which assesses students' receptive vocabulary, is a commonly used vocabulary measure, we tested expressive vocabulary rather than receptive vocabulary because K-PAVE is intended to give students many opportunities to practice their expressive skills in addition to receptive word learning. K-PAVE instructional practices encourage students' oral language use during Interactive Book Reading and extended Adult-Child Conversations. In addition, there was evidence from the previous quasi-experimental evaluation of the preschool PAVE intervention of statistically significant positive effects on students' expressive vocabulary. Furthermore, the National Early Literacy Panel (NELP 2009) found an average correlation, across 30 randomized studies between expressive language comprehension outcomes and later reading performance of .48. Measures of receptive vocabulary had relatively weak relationships with both decoding and reading comprehension. Consequently, we concluded that a measure of

[22] Baseline testing was delayed for 3.9% of students in control schools because parental permissions, although collected by the school earlier, were provided to the study team late.

expressive vocabulary was better aligned with the immediate, primary goals of the program, as well as with the secondary goal of an extended impact on later reading comprehension.

The recently updated version of the EVT has been co-normed and standardized on a representative sample of American children with attention to gender, race/ethnicity, geographic region, socioeconomic status, and special education needs. The EVT–2 has been used in criterion studies with other language assessments such as the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk 1999) and Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001). The EVT–2 has internal consistency (split half) reliabilities of .94–.95 and has a test-retest reliability of .95 (Williams 2007). Correlations between the EVT–2 and other tests are .84 for the Peabody Picture Vocabulary Test–4 (Dunn & Dunn 2007) for children ages 5–6; .68–.80 for the receptive language, expressive language, and core language scales on the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4; Semel, Wiig, & Secord, 2003) for children ages 5–8; and .59–.76 for the GRADE total reading score in kindergarten.

The EVT–2 raw scores were used to calculate a standard score using student age. The standard scores allow the performance of students in intervention classrooms to be compared with that of a national sample of children of the same age. Such comparisons can be used to address the policy implications of any changes in the achievement gap between the study sample of at-risk children and a national sample.

Academic knowledge. The study also measured general or background knowledge, a secondary outcome, using the Woodcock-Johnson III/Normative Update (WJ–III/NU) Academic Knowledge subtest (Woodcock, McGrew, Schrank, & Mather 2007). The Academic Knowledge subtest is a suggested outcome measure for interventions that provide a language-rich environment, frequent exposure to words, reading aloud to children, and text talk (Wendling, Schrank, & Schmidt 2007). The test focuses on background knowledge in science, social studies, and humanities:

- Science items include naming or pointing to body parts and pictures of animals and vegetables and answering questions such as, “What does a tadpole, or polliwog, become when it grows up?”
- Social studies items include naming or pointing to items of clothing, furniture, grooming, tools, and buildings and answering questions such as, “What is the person called who fixes teeth?”
- Humanities items include naming or pointing to arts and crafts supplies, colors, and musical instruments and completing the phrase, “Once upon a ….” or answering, “What is a large group of singers called?”

A single item response theory–based scale score (*W*-score) was calculated based on a nationally representative norm group. Reported technical characteristics of the WJ–III/NU Academic Knowledge test indicate that internal consistency (split half) reliability is .92 for students age 5 and .82 for children age 6 (Woodcock, McGrew, Schrank, & Mather 2007).

Listening comprehension. The study also measured students’ comprehension, since increasing children’s vocabulary and knowledge about the world is a pathway to stronger skills

in comprehending spoken language and print. For kindergarten students, most of whom have not yet learned to read connected text, comprehension is most validly measured through listening to speech rather than through reading. Listening comprehension at the end of kindergarten was measured using the Kaufman Test of Educational Achievement–II (KTEA–II) Listening Comprehension subtest (Kaufman & Kaufman 2004). This test assesses listening ability and understanding, without assessing reasoning or memory for details. The assessment involves listening to short passages read orally and answering comprehension questions. Student age was used in calculating a single standardized score based on a nationally representative norm group. Listening comprehension is a secondary outcome in the impact analysis.

Reported technical characteristics of the subtest indicate that the internal consistency (split-half) reliability for this measure in kindergarten is .84 (Kaufman & Kaufman 2004). Results from a confirmatory factor analysis with students in grade 1 and higher indicate that the correlation between the oral language factor (listening comprehension and oral language subtests) and the reading factor is .91 and that the errors for the listening comprehension subtest and the reading comprehension subtest are correlated. The correlation in grades 1–5 between the KTEA–II Listening Comprehension subtest and the WJ–III Listening Comprehension test is .77.

Observations of instruction in intervention and control classrooms. Teachers' instructional behavior was assessed through direct classroom observation at baseline and at the end of the school year. Half-day observations focused on vocabulary and literacy instruction. Baseline observations were completed before the K-PAVE intervention training. Posttest observations were timed for weeks 22–24 of the 24-week intervention period.[23] Observations were conducted by trained members of the evaluation team who were independent of the K-PAVE intervention and unaware of the school assignment to intervention or control conditions.

Classroom observation measures. Three classroom observation measures were administered and are discussed below (see Table 2.6 for a summary and Appendix F for detailed descriptions).

- *Classroom Assessment Scoring System.* The Classroom Assessment Scoring System (CLASS K–3; Pianta, La Paro, & Hamre 2008) assesses the quality of interactions and instruction and yields ratings on instructional support, emotional support, classroom organization, and a combined global quality rating.

  The CLASS, a time-sampling observation tool, was based on at least two hours of classroom observation. Coding involved 30-minute cycles, during which 10 dimensions of classroom instruction were rated on a seven-point Likert scale. Observers also documented the academic content covered (vocabulary/comprehension, other language arts, or other academic content).

  Two of the three CLASS domains—instructional support and emotional support—more closely reflect the goals of the K-PAVE intervention. K-PAVE trains teachers on techniques for vocabulary instruction, Interactive Book Reading, and informal conversations with students, techniques that can be expected to affect dimensions of instructional support (concept development, quality of feedback, and language modeling). The emphasis on

[23] The initial K-PAVE intervention training workshop was staggered over four weeks, so not all schools began the intervention at the same time. For this reason, weeks 22–24 of the intervention period spanned a five-week period, depending on when the intervention training was delivered.

informal conversations about topics of students' interest and choosing may be expected to influence emotional support (positive classroom climate, teachers' regard for student perspectives, and teacher sensitivity). Therefore, scores for these two domains were used as secondary outcomes for the study. In addition, data from the CLASS observation on academic content were used to describe the distribution of classroom time across content areas. This variable was used to examine whether implementation of the K-PAVE supplementary program reduced the time spent on other literacy-related instruction.

- *Vocabulary Record.* The Vocabulary Record, which documents the vocabulary support provided by teachers during a book read-aloud and during each CLASS cycle, yields two measures: the number of words defined by the teacher or assistant teacher during book reading and the average number of words defined during other instructional time. The Vocabulary Record is coded throughout the entire classroom observation; the observer documents every word that the teacher or assistant teacher introduces—by providing a definition or synonym, illustrating with a picture, providing a contrast, or asking a student for word meaning—during the book reading when the Read Aloud Profile–Kindergarten is being coded and during each CLASS cycle. The total number or words documented during the read-aloud was tallied for a measure of the *number of words introduced during the book reading*. The total number of words documented during each CLASS cycle was tallied and averaged (i.e., divided by the number of CLASS cycles) for a measure of the *average number of words introduced during other instructional time*. The Vocabulary Record created for this study is an adaptation of the vocabulary component of the Instructional Practice in Reading Inventory (IPRI; Smith et al. 2005).
- *Read Aloud Profile–Kindergarten.* The Read Aloud Profile–Kindergarten (RAP–K) documents teachers' comprehension support and questions while reading aloud to students. Comprehension supports include providing background information, making connections to students' experiences, and asking concrete or factual questions to clarify meaning. Higher order questions include questions asking students to analyze, explain, predict, imagine, make inferences, or generate hypotheses. Two measures were generated: *number of comprehension supports* and *number of higher order questions*. When teachers did not read aloud to students during the classroom observation, the reading behaviors measured by the RAP-K were coded as not occurring (i.e., occurring zero times) rather than as missing. The RAP-K, created for this study, is an adaptation of the Read Aloud Profile (RAP) instrument from the Observation Measures of Language and Literacy Instruction (OMLIT; Goodson, Layzer, Smith, & Rimdzius 2004).

Vocabulary and comprehension support composite created from RAP-K and Vocabulary Record variables. The two variables created from the Vocabulary Record—the number of words introduced during book reading and the average number of words introduced during other instructional time—and the two variables created from the RAP-K—number of comprehension supports and number of higher order questions—are all measures of the broader construct of *vocabulary and comprehension support*. The K-PAVE intervention is intended to improve teachers' vocabulary and comprehension instructional behaviors and thereby improve students' vocabulary development. In order to test the impact of K-PAVE on the vocabulary and comprehension support construct overall rather than test its impact on the specific behaviors that comprise the construct, we created a single composite measure from the four variables. (Creation of the composite is described in Appendix F.)

In addition to combining the four variables into a single composite for substantive reasons, doing so also has the benefit of minimizing the number of hypothesis tests being conducted. The risk of type I error (i.e., false positives) when we test the impact of K-PAVE on a single composite measure of vocabulary and comprehension support is lower than if we were to test the impact of K-PAVE on each of the four variables that comprise the construct.

Table 2.6 Classroom/teacher measure and outcome variables

Study area	Measure	Outcome variable	Status in analysis
Instructional support	Classroom Assessment Scoring System (CLASS)	Overall rating on instructional support (1–7)	Secondary
Emotional support	CLASS	Overall rating on emotional support (1–7)	Secondary
Amount of literacy-related instruction in areas other than vocabulary and comprehension	CLASS	Proportion of CLASS cycles in which literacy-related content other than vocabulary, comprehension, and oral language is covered	Secondary
Vocabulary and comprehension support	• Vocabulary Record • Read Aloud Profile–Kindergarten	Vocabulary and comprehension support composite: • Number of words teacher/assistant introduces or asks students to define during book read-aloud • Average number of words teacher or assistant teacher introduces or asks students to define during other instructional time • Number of comprehension supports during reading • Number of higher order questions during reading	Secondary

Covariate measures of student, teacher, and school characteristics. Data were collected on the characteristics of students, teachers, and schools. Data on student demographics and school characteristics were included as covariates in the analysis of impacts on students, and data on school characteristics and teacher background were included as covariates in the analysis of impacts on classroom instruction.

Student demographic characteristics were collected from district administrative records on date of birth, gender, race/ethnicity, special education status, status as an English language learner, retention in kindergarten, eligibility for free or reduced-price meals, and whether attended preschool.

For teachers, baseline data were collected using a short (approximately 10 minutes) survey on race/ethnicity, age, training/education, and number of years of teaching experience and of teaching kindergarten. (The teacher survey is presented in Appendix G.) The survey was completed by teachers in both intervention and control classrooms as part of the baseline classroom observation visit (when the CLASS, RAP-K, and Vocabulary Record instruments

were collected, as described above). Observers gave teachers the paper-and-pencil survey to complete during the classroom visit.

Data on school characteristics gathered from the Mississippi Department of Education, districts, and schools included the percentage of students in the school receiving free or reduced-price meals, the school's Achievement Level Index[24] (a score based on student performance in grades 3 and higher on the Mississippi Curriculum Test in 2006/07), school racial/ethnic composition, and the literacy curricula used in kindergarten classrooms.

Fidelity of implementation. Fidelity of implementation was based on two factors: the degree to which intervention teachers participated in the K-PAVE teacher training and support, and whether or not intervention teachers implemented the 12 K-PAVE instructional strategies as intended. The participation in training and support was based on attendance records, while fidelity of implementation was based on classroom observations by members of the training team.

Fidelity of implementation was based on observation of intervention classrooms in late February or early March (corresponding to weeks 17–20 of the 24-week intervention period), after teachers had the opportunity to participate in all of the planned K-PAVE training and support. The measure of fidelity, the Training Fidelity Checklist, was provided by the K-PAVE developer (see Appendix H for the K-PAVE Training Fidelity Checklist and coding instructions). Data collectors trained by the developer scored fidelity of implementation in each intervention classroom based on a 90-minute observation. Teachers were assigned a fidelity score ranging from 1 to 12, indicating the number of strategies to which they demonstrated fidelity. Based on the K-PAVE developer's guidelines about the number of strategies implemented with fidelity, teachers were then assigned a fidelity rating. (Fidelity ratings are discussed in further detail in Chapter 3.)

Observers also kept detailed field notes to indicate the evidence they used to score implementation. Inter-rater agreement on the fidelity scoring was not assessed during the training. However, as part of the fidelity observation, observers were trained to make detailed field notes, describing the evidence on which they based their fidelity scoring. A member of the developer's team, who did not observe the teacher directly, subsequently scored the fidelity of implementation based solely on the field notes. There was a high level of agreement between the fidelity scores assigned by the observers and by the independent scoring based on the field notes (Cohen's Kappa = .98 across the 12 individual instructional strategies).

[24] A school's Achievement Level Index is created based on student performance on the Mississippi state accountability test (Mississippi Curriculum Test, MCT), which is administered to all students in grade 3 and higher. The Achievement Level Index (ALI) corresponds to the School Performance Classification (described above); the ALI is measured on a continuous scale ranging from 100 to 600, while School Performance Classification is a categorical rating (i.e., low performing, underperforming, successful, exemplary, and superior). The ALI score is created based on the percentage of students in the school who scored basic or higher on the MCT and the percentage of students in the school who scored proficient or higher on the MCT. Schools with scores in the 100 range are rated as having a School Performance Classification of "low performing"; schools with scores in the 200 range are rated "underperforming"; schools with scores in the 300 range are rated "successful"; schools with scores in the 400 range are rated "exemplary"; and schools with scores in the 500 range are rated "superior."

Training of data collectors to ensure reliable data

The study required two types of data collectors—student assessors and classroom observers. Their recruitment and training are described below.

Qualifications of data collectors. Data collectors were recruited from local universities and community colleges in Mississippi and through contacts at the Mississippi Department of Education. Data collectors included students with a background in education and retired teachers, school counselors, and school administrators. Data collectors were independent of the intervention implementation and unaware of intervention status. Data collection procedures were identical in intervention and control schools, since data collectors did not know schools' intervention status. (Student assessment and classroom observation procedures are described in Appendix I.)

Training data collectors to level of research reliability. Each group of data collectors—student assessors and classroom observers—attended one week of training and had to pass reliability testing before being hired to collect data. Child assessors were trained by senior Abt Associates staff experienced with the child assessments used in this study. Trainees received thorough instruction on test administration and scoring rules and had numerous opportunities to practice mock test administrations.[25] Classroom observers were trained to use the CLASS instrument by staff from the University of Virginia Center for the Advanced Study of Teaching and Learning (CASTL), where the instrument was developed. Observers were trained to administer the RAP-K and Vocabulary Record instruments by senior Abt Associates staff involved in the development of or with extensive experience using the instruments. Training on all observation tools involved thorough instruction on coding rules, illustration of codes using video examples of classroom instruction, and numerous opportunities to practice coding video segments.

Criteria developed before training were used to determine whether trainees had met the required standards. Student assessors were required to conduct a mock administration of each student assessment without any major administration errors (such as scoring errors or incorrectly establishing basal or ceiling criteria) and with fewer than two minor errors (such as neglecting to point at a picture when reading a prompt). Classroom observers were required to code five video segments of classroom interactions for the CLASS instrument. For the RAP–K and the Vocabulary Record, classroom observers were required to code two video segments of classroom read-alouds. All videotapes had been coded in advance by master coders, and observers were required to have at least 80% agreement for all coded video segments. The

[25] For all child assessments—Expressive Vocabulary Test-2; Woodcock-Johnson III/NU Academic Knowledge test; and Kaufman Tests of Educational Achievement-II Listening Comprehension test—assessors were trained to follow administration and scoring guidelines outlined by the test publisher. A substantial amount of training time was devoted to learning criteria outlined in test manuals for judging student responses as correct, incorrect, or requiring further prompting. In mock administrations of each test, trainers provided increasingly complex responses to test items, to give trainees experience scoring a range of responses.

average rate of agreement with the master codes was 82% on the CLASS,[26] 88% on the RAP–K,[27] and 85% on the Vocabulary Record.[28]

For student assessment, 21 of 34 trainees (62%) were certified to collect data, and for baseline data collection, 10 of 14 trainees (71%) were certified as reliable. For posttest data collection, both returning and new data collectors were trained and certified. Among student assessors, all 13 returning assessors (100%) and six of nine new trainees (67%) were certified to collect data. Among classroom observers, all four returning observers (100%) and five of eight new trainees (63%) were certified as reliable.

In addition, data collectors received close oversight during the first weeks of data collection to identify any problems before extensive data were collected. Experienced data collectors accompanied new ones on early data collection visits to provide guidance and answer questions. Within the first week of data collection, trainers held conference calls with data collectors to discuss questions and challenges. No measure of interrater reliability was collected during these visits.

Quality assurance process for data. Student assessments and classroom observation data underwent a thorough quality assurance protocol (described in Appendix J).

Response rates for data collection

As shown in Table 2.7, the study achieved high response rates at all levels—schools, classrooms, teachers, and students.

School-level and classroom-level attrition. During the study one intervention school dropped out; no data were collected on student demographics or on posttest classroom instruction or student outcomes. Without posttest data or student-level covariates, the school could not be included in the analysis. The school that dropped out of the study has characteristics similar to the average school in the intervention group and the average school in the sample overall.[29] The school was in the block of intervention schools with either a local reading program or no reading program. The remaining schools in the block were weighted to adjust for the loss of the school from this block (see Appendix K for a discussion of the weighting used), and sensitivity models were estimated without weights to examine sensitivity to weighting (see Appendix N).

Because one school dropped out of the study, two classrooms and 23 students were lost from the intervention group. A total of 64 schools (30 intervention and 34 control), 128 classrooms (60 intervention and 68 control), and 1,296 students (596 intervention and 700 control) remained in the analytic sample. Appendix L illustrates the attrition from data collection

---

[26] Each observer's average rate of agreement with master codes across all five coded CLASS video segments ranged from 80% to 84%.

[27] Each observer's average rate of agreement with master codes across both coded RAP-K video segments ranged from 80% to 96%.

[28] All observers were within one word of the number of words identified for the master coders on each of the two coded video-recordings. The rate of agreement ranged from 63% to 100%. The low incidence of new vocabulary words being introduced contributed to a lower than 80% rate of agreement for three classroom observers.

[29] To maintain confidentiality, we do not report specific characteristics of the school that dropped from the study in comparison to the remaining schools.

at the school, classroom, and student levels. There was no other attrition at the school or classroom levels.

Student-level attrition. The flow of students through the study—including recruitment, random selection, and attrition from data collection—is shown in Figure 2.3 for both intervention and control groups. At posttest, student attrition was 8.1% in the intervention group and 5.9% in the control group. The higher attrition rate in the intervention group reflects the loss of students from the school that dropped out of the study. Excluding the students from that school brings the attrition rate in the intervention group down to 4.5%. There is no statistically significant difference between these two rates ($\chi^2 = 1.14$, $p = .29$) or between the two rates when students from the school that dropped out are included ($\chi^2 = 2.19$, $p = .14$). With the exception of the one intervention school that dropped out of the study, the reasons for attrition were that students moved out of state, transferred to a nonstudy school in Mississippi, and were absent during the data collectors' visits. (See section later in the chapter on data analysis methods and Appendix M for details on the imputation of missing data. Appendix N reports sensitivity analyses examining the influence of missing data imputation on impact estimates and their standard errors.)

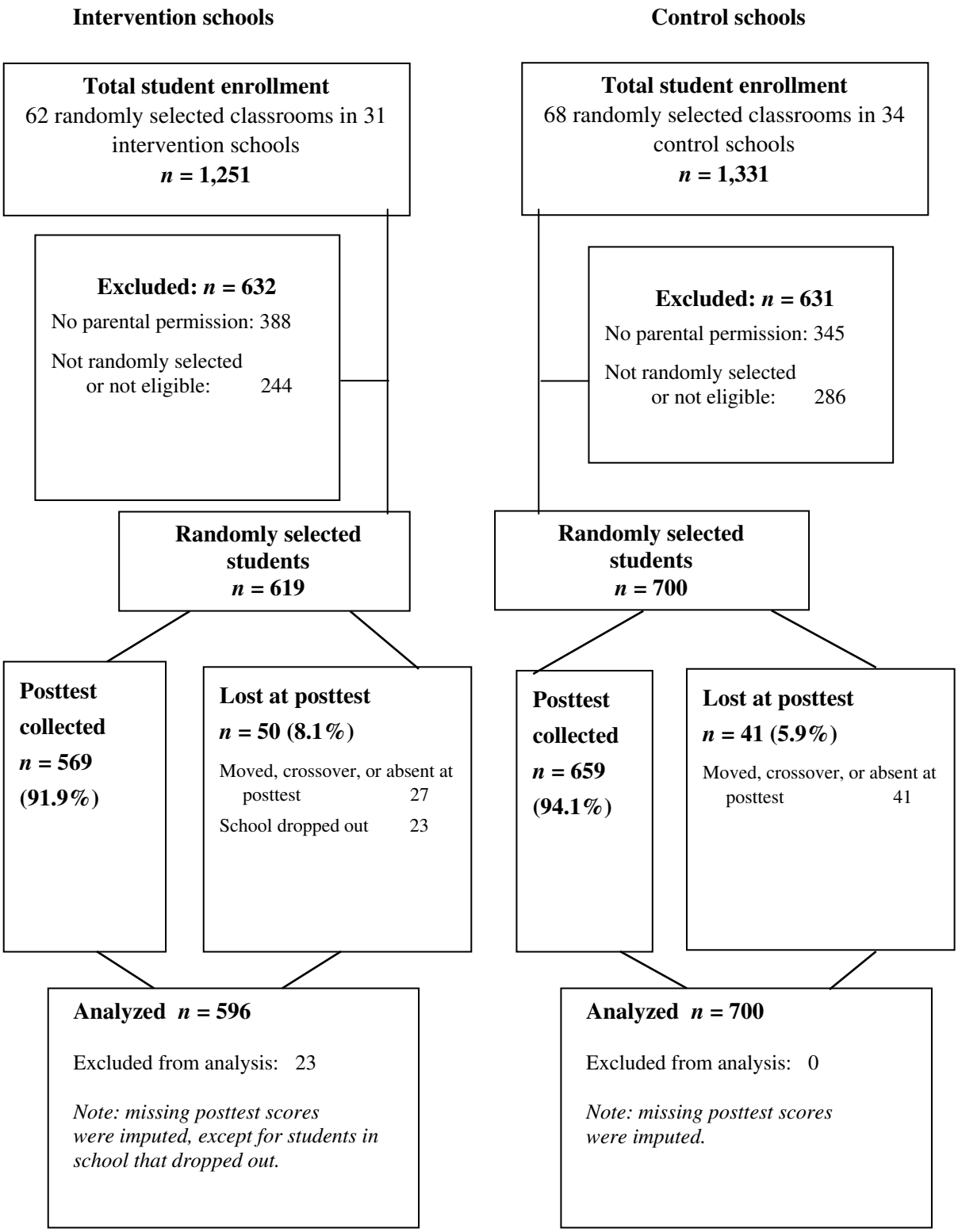
Table 2.7 Response rates for data collected from schools, classrooms, teachers, and students for intervention and control groups (percent)

	Baseline response rate			Posttest response rate		
Level and variable	**Overall**	**Intervention**	**Control**	**Overall**	**Intervention**	**Control**
Schools (number)	65	31	34	65	31	34
Demographics	100	100	100	na	na	na
Classrooms (number)	130	62	68	130	62	68
Reading program	100	100	100	na[a]	na	na
Classroom Assessment Sorting System	100	100	100	98.5	96.8	100
Read Aloud Program–Kindergarten	100	100	100	97.7	96.8	98.5
Vocabulary Record	100	100	100	98.5	96.8	100
Teachers (number)	130	62	68	130	62	68
Teacher demographic survey	99.2	98.4	100	na	na	na
Language sample	93.8	88.7	98.5	93.8	93.4	94.1
Students (number)	1,319	619	700	1,319	619	700
Demographics	97.5	95.5	99.9	na	na	na
Expressive vocabulary (EVT–2)	96.3	96.0	96.6	93.5	92.2	94.6
Listening comprehension (KTEA–II)	94.4	93.7	95.0	93.4	92.2	94.4
Academic knowledge (WJ–III/NU)	95.1	93.9	96.1	94.3	92.1	94.6
Passage comprehension (WJ–III/NU)	96.3	96.0	96.6	93.5	92.2	94.6

na is not applicable because the variable was collected only once during the school year, rather than at baseline and posttest.

Note: One school (two classrooms) dropped out of the study after baseline data collection, which means all posttest measures and student demographic data are missing for one intervention school, two intervention classrooms/teachers, and 23 intervention students.

Figure 2.3 Flow of students through the study



Nonparticipants and crossovers

Students. Forty-five students were categorized as nonparticipants—students who moved out of state or transferred to a nonstudy school in Mississippi either before baseline testing or between pretest and posttest and therefore have no posttest data. Of the 45 students, 21 were in intervention schools and 24 in control schools (Table 2.8). To maintain the statistical equivalence of the two control groups, all students—both participants and nonparticipants—were included in the analysis (except students in the school that dropped out of the study). To retain nonparticipants in the analysis, values were imputed for missing pretest and posttest data (see Appendix M on missing data imputation). The inclusion of nonparticipants in the analysis could lead to underestimation of the effect of the K-PAVE intervention, depending on the number of nonparticipants. Sensitivity analyses examining the influence of nonparticipants on impact estimates and standard errors are reported in Appendix N.

Table 2.8 Nonparticipants in intervention and control groups

Condition	Intervention (n = 596)	Control (n = 700)	Test of difference
Transferred out of state or to a nonstudy school (nonparticipants)	21 (3.4%)	24 (3.4%)	
Tested at pretest only	10 (1.6%)	18 (2.6%)	$\chi^2 = 3.56$
Not tested at pretest or posttest	11 (1.8%)	6 (0.9%)	$p = .06$

There were only five crossover students—students who were initially enrolled in a control school but switched to an intervention school or students who were initially enrolled in an intervention school but switched to a control school during the study. The crossover rate was not significantly different for the intervention and control schools (p=.99).[30] If there are many crossovers, this kind of movement can compromise the integrity of the impact estimates. The low rate of crossovers in this study is unlikely to threaten the validity of the estimates. To preserve the integrity of the random assignment, crossover students were analyzed in their original assigned condition. This could bias the analysis against finding significant effects of the K-PAVE intervention. (See Appendix N for sensitivity analyses examining the influence of student crossovers.)

Teachers. No intervention teachers left midyear. Because there was no teacher turnover in the intervention schools, there was no need to offer the K-PAVE workshop training to midyear replacement teachers.

ANALYTIC SAMPLE

Because one intervention school dropped out of the study, the analytic sample included one fewer school, two fewer classrooms, and 23 fewer students than the initial sample after randomization. The loss of that one school resulted in an overall attrition rate of 2% and a

[30] Fischer's exact test was used to test for differences in crossovers between intervention and control groups because of small expected values for cells.

differential attrition rate of 3%.[31] The analytic sample had 64 schools (30 intervention and 34 control), 128 classrooms (60 intervention and 68 control), and 1,296 students (596 intervention and 700 control). Table 2.9 compares the initial sample and the analytic sample.

Table 2.9 Comparison of initial sample at randomization and final analytic sample

Level and condition	Initial sample at randomization	Final analytic sample
Schools	65	64
Intervention	31	30
Control	34	34
Classrooms	130	128
Intervention	62	60
Control	68	68
Students	1,319	1,296
Intervention	619	596
Control	700	700

Characteristics of teachers and students in intervention and control schools

There are no statistically significant differences in characteristics between teachers in the intervention and control groups (Table 2.10) or between students in the intervention and control groups (Table 2.11). Nonetheless, to obtain more precise estimates of K-PAVE impacts on classroom instruction, all teacher characteristics were included as covariates in the analysis of impacts on classroom instruction because teacher characteristics might relate to classroom instructional practices. Similarly, to obtain more precise estimates of K-PAVE impacts on students, all student characteristics were included as covariates in the analysis of impacts on students because these characteristics are likely related to student performance.

[31] The loss of 1 of 31 intervention schools gives an attrition rate in the intervention group of 3%. The attrition rate in the control group is 0% so the differential attrition rate is 3% (intervention group attrition rate) minus 0% (control group attrition rate).

Table 2.10 Characteristics of teachers in the analytic sample, by intervention condition

Characteristic	Intervention[a] (60 classrooms)	Control (68 classrooms)	Overall (128 classrooms)	Test of difference[b]
Race				
African American	40.7%	48.5%	44.9%	$t = –0.81$, $p = .42$
White	59.3%	51.5%	55.1%	
Education				$t = 0.03$, $p = .98$
College	39.4%	36.8%	38.1%	
Some graduate courses	22.4%	27.9%	25.4%	
Graduate degree	37.9%	35.3%	36.5%	
Certifications				
Early childhood	71.7%	79.4%	75.8%	$t = –0.96$, $p = .34$
Reading	11.7%	13.2%	12.5%	$t = –0.30$, $p = .76$
Years teaching				$t = 1.05$, $p = .30$
Mean	17.0 years	14.6 years	15.7 years	
Standard deviation	12.3 years	11.3 years	11.8 years	
Years teaching kindergarten				$t = 0.91$, $p =. 37$
Mean	10.9 years	9.4 years	10.1 years	
Standard deviation	9.1 years	8.7 years	8.9 years	

a. Sample includes all intervention classrooms ($n = 60$) and control classrooms ($n = 68$). Distributions of teacher characteristics are assumed to be the same for cases with missing data as for cases with non-missing data. Rates of missing data range from 0.0% to 3.3%.

b. A two-level model was used to test for baseline differences between intervention and control groups on teacher characteristics at the school level to account for the nesting of teachers within schools. The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates. A linear model was used to test for differences in dichotomous and ordinal variables—gender, race, education, early childhood certification (yes/no), and reading certification (yes/no)—to account for the nested data structure. Thus, *t*-tests rather than chi-square tests were conducted.

Table 2.11 Characteristics of students in the analytic sample, by intervention condition

Characteristic	Intervention (n = 596)	Control (n = 700)	Overall (n = 1,296)	Test of difference[a]
Gender				
Female	49.2%	50.6%	50.0%	t = –0.47, p = .64
Male	50.8%	49.4%	50.0%	
Race				
African American	87.6%	82.8%	85.0%	t = 0.81, p = .42
White	11.5%	14.8%	13.3%	
Other	0.8%	2.4%	1.7%	
Eligibility for free or reduced-price meals				
Yes	93.4%	92.5%	92.9%	t = 0.26, p = .80
No	6.6%	7.5%	7.1%	
Has an Individualized Education Program				
Yes	8.8%	7.7%	8.2%	t = 0.55, p = .58
No	91.2%	92.3%	91.8%	
Age at posttest				
Mean	6 yrs, 1.9 mo	6 yrs, 1.8 mo	6 yrs, 1.9 mo	
Standard deviation	4.8 mo	4.6 mo	4.7 mo	

Note: The distribution of student characteristics is assumed to be the same among missing data as among nonmissing data. Rates of missing data range from 0.1% to 1.8%.

a. A three-level model was used to test for baseline differences between the intervention and control groups in student characteristics at the school level to account for the nesting of students within classrooms and classrooms within schools. The test of baseline differences was adjusted for the multilevel structure of the data but was not adjusted for covariates. A linear model was used to test for differences in dichotomous variables—gender, race, eligibility for free or reduced-price meals, and having an Individualized Education Program—to account for the nested data structure. Thus, *t*-tests rather than chi-square tests were conducted.

Student outcomes and classroom instruction at baseline for intervention and control schools

Table 2.12 shows the baseline pretest scores on student outcome measures for intervention and control groups. Before the start of the K-PAVE intervention, students were administered standardized assessments of expressive vocabulary (EVT–2), academic knowledge (WJ–III/NU Academic Knowledge subtest), and listening comprehension (KTEA–II Listening Comprehension subtest). Scores for each assessment were standardized based on a norming sample and have a mean of 100 points and a standard deviation of 15 points. In both intervention and control groups, students in the sample have mean baseline standard scores approximately two-thirds of a standard deviation below the norm for their age on expressive vocabulary and academic knowledge and slightly lower for listening comprehension. There were no statistically significant differences at baseline between intervention and control groups on any of the student outcomes.

Table 2.12 Baseline pretest scores on student outcomes for intervention and control groups

	Unadjusted pretest means		Test of baseline differences in student outcomes[a]	
Student outcome measure	Intervention group mean (standard deviation) *n*	Control group mean (standard deviation) *n*	Estimated difference (standard error) *n*	Test of difference (*p*-value)
Expressive vocabulary	91.1 (12.3) 574	90.7 (11.4) 676	0.40 (1.10) 1250	.72
Academic knowledge	90.5 (11.0) 574	90.6 (11.3) 675	–0.05 (1.05) 1,249	.96
Listening comprehension	88.2 (12.8) 560	87.0 (13.4) 665	1.16 (1.52) 1,225	.44

Note: The sample includes 596 intervention group students in 60 classrooms in 30 schools and 700 control group students in 68 classrooms in 34 schools.
a. A three-level model, with student, classroom, and school levels, was used to test for the baseline difference between intervention and control group means in student outcome scores at the school level. The model had the same multilevel structure as the model used to test K-PAVE impacts on student outcomes at posttest. (See Appendix K for impact model specifications.) The test of baseline differences was adjusted for the multilevel structure of the data but was not adjusted for covariates.

Table 2.13 shows the baseline classroom instruction measures for intervention and control groups. The average rating of instructional support of approximately 2.6 points in both intervention and control classrooms indicates quality at baseline at the high end of low quality.[32] The average rating of emotional support of 5.2 points in intervention groups and 5.1 points in control groups indicates quality at baseline at the high end of moderate quality.

Intervention and control groups do not differ significantly at baseline on these two classroom instruction measures or on the measure of vocabulary and comprehension support. However, the percentage of CLASS cycles observed at baseline during which the teacher spent time on nonvocabulary literacy instruction (such as print concepts, phonological awareness, and alphabet knowledge) was 10 percentage points higher in intervention schools than in control schools ($t = 2.08$, $p = .04$). This difference is assumed to occur by chance rather than as the result of any bias in the random assignment of schools or the random selection of classrooms for data collection. Because there were no missing data on the classroom observation measures, the difference is not the result of nonresponse bias. The analysis of the impact of K-PAVE on instructional time included the baseline measure for each outcome to increase the precision of the impact estimate. In this case, inclusion of the baseline measure in the impact analysis also reduces variability in the impact estimates that can result from chance nonequivalence between the intervention and the control groups.

[32] The Classroom Assessment Scoring System (Pianta et al. 2008), used to rate instructional support and emotional support, defines ratings of 1 and 2 as low quality, ratings of 3–5 as moderate quality, and ratings of 6–7 as high quality.

Table 2.13 Baseline measures of classroom instruction for intervention and control groups (*n* = 64 schools, 128 classrooms)

	Raw pretest means		Test of baseline differences in classroom instruction outcomes[a]	
Classroom instruction measure	Intervention group mean (standard deviation)	Control group mean (standard deviation)	Estimated difference (standard error)	Test of difference (*p*-value)
Vocabulary and comprehension support	–0.002 (1.00)	0.002 (1.01)	–0.007 (.183)	.97
Instructional support	2.62 (0.73)	2.67 (0.69)	–0.049 (0.156)	.75
Emotional support	5.21 (.67)	5.06 (.81)	0.152 (.133)	.26
Instructional time on literacy areas other than vocabulary and comprehension	.56 (.23)	.46 (.24)	.103* (.049)	.04

* $p < .05$.

Note. The sample includes all intervention classrooms (60 classes in 30 schools) and all control classrooms (68 classrooms in 34 schools). Vocabulary and comprehension support is measured as a *z*-score with a mean of 0 and a standard deviation of 1. Instructional support and emotional support are measured on a seven-point Likert scale. Instructional time on other literacy areas is measured as a proportion, with a 0–1 possible range.

a. A two-level model, with classroom and school levels, was used to test for the baseline difference between intervention and control group means on measures of classroom instruction at the school level. The model had the same multilevel structure as the model used to test K-PAVE impacts on classroom instruction at posttest. (See Appendix K for impact model specifications.) The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates.

DATA ANALYSIS METHODS

This section describes the analytic approach used to address research questions on the impact of the K-PAVE intervention primarily on kindergarten students' expressive vocabulary and secondarily on kindergarten students' academic knowledge and listening comprehension. It also describes the analytic approach used to address research questions on classroom instruction in vocabulary and comprehension support, instructional support, and emotional support and the extent to which aspects of literacy other than vocabulary and comprehension received instructional focus.

Based on the presence of a positive impact of K-PAVE on students' expressive vocabulary in kindergarten, a subsequent report will examine the impacts of exposure to K-PAVE in kindergarten on students' expressive vocabulary, academic knowledge, listening comprehension, and passage comprehension in grade 1. A separate exploratory report will investigate impacts of K-PAVE on kindergarten students' lexical diversity and variation in impacts on students in kindergarten and grade 1 depending on student characteristics (such as age, gender, and entry-level pretest scores). The exploratory report will also investigate impacts on kindergarten teachers' lexical diversity and other instructional practices.

Impacts on students

A three-level hierarchical linear model (HLM), with school, classroom, and student levels, was used to model the impact of the K-PAVE intervention on students (the model is

presented in Appendix K). The model provides an estimate of the average impact of the intervention on students across all schools at a given time (at the end of kindergarten) and an estimate of this impact's standard error. The HLM is appropriate for this analysis because the study used a multilevel design with students nested within classrooms and classrooms within schools. The multilevel modeling also parses the variance among students, classrooms, and schools to produce more accurate estimated standard errors (Raudenbush & Bryk 2002).

The student outcome variable (EVT–2 posttest score), student baseline test score, and student covariates were included in the model at the student level, with residual variation between students within a classroom also represented. Variation between classrooms within a school in the average student outcome score was modeled at the classroom level. Included at the school level were school covariates and an indicator variable indicating whether a school is in the intervention group or the control group; this level also models residual variation between schools. The parameter for the intervention variable indicates the impact of the K-PAVE intervention on the specified student outcome. A *t*-test with a .05-level of significance as the criterion was used to test the null hypothesis that the intervention effect is zero. A positive and statistically significant parameter estimate means that the K-PAVE intervention improves the student outcome by the magnitude of the parameter estimate. A standardized effect size was calculated by dividing the estimated impact from the model by the standard deviation of the outcome variable in the control group. The control group standard deviation was used, as recommended in Burghardt et al. (2009), because the intervention could affect the standard deviation in the intervention group.[33]

Student-level covariates used in the analysis included (see Appendix O for definitions):

- Score on the student outcome measure at baseline (pretest score).
- Gender.
- Race.
- Eligibility for free or reduced-price meals.
- Having an Individualized Education Program.

All covariates were included in the model when analyzing each student outcome variable. However, for each student outcome measure, only the corresponding pretest score for that measure was included.

School-level covariates used in the analysis included:

- Reading initiatives (Reading First, a Mississippi state reading initiative, or other).
- Achievement Level Index.
- Percentage of students who are African American.

[33] Our view is that standardized effect sizes should be used to help interpret the magnitude of impacts. However, we believe that tests of statistical significance should be conducted on impact estimates measured in nominal units rather than on the standardized effect sizes. Therefore, we do not conduct tests of statistical significance on the standardized effect sizes.

- Percentage of students eligible for free or reduced-price meals.
- Locale (rural, small town, or large town/fringe of city).
- Location in or out of the Mississippi Delta region.

All school covariates were included when analyzing each student outcome variable.

Impacts on classrooms

The impacts of the K-PAVE intervention on instructional practices, controlling for teacher and school characteristics, were estimated using a multilevel model to account for the clustering of two classrooms per school. The model included a classroom level (level 1) and a school level (level 2). Because of the limited degrees of freedom at the classroom level (due to sampling only two classrooms per school), teacher characteristics were controlled for at the school level. The average value for the school was calculated for each teacher characteristic. The multilevel model used to test for K-PAVE impacts on classroom instruction is specified in Appendix K.

The same school-level covariates included in the models estimating impacts on students were included in the models estimating impacts on classroom instruction.

The following teacher characteristics were included in the analysis as covariates in the school-level model (and are defined in Appendix O):[34]

- Gender.
- Race.
- Highest level of education.
- Number of years teaching.
- Number of years teaching kindergarten.
- Having teaching certification in early childhood.
- Having teaching certification in reading.

A dummy variable indicating whether a school was assigned to the intervention or control group was included at the school level. The parameter estimate for the intervention indicator estimates the impact of K-PAVE on the specified classroom instruction outcome. A positive and statistically significant parameter estimate indicates that K-PAVE affects instruction in the desired direction. The magnitude of the parameter indicates the estimated magnitude of the average impact. Effect sizes of classroom impacts were calculated following the approach described above for student impacts.

[34] There were plans to include covariates for teacher ethnicity (Hispanic or not) and certification in special education, but no teachers were Hispanic, and none was certified in special education.

Adjustments for multiple comparisons

In testing for an impact on an outcome, there is some chance that the null hypothesis of no impact will be rejected when there is no true impact (type I error). To limit the likelihood of such false positives to an acceptable level, the criterion for rejecting the null hypothesis was set to $p = .05$ for a single hypothesis test. However, the chance of one or more false positives increases as tests of impacts are conducted on more outcomes. To reduce the heightened chance of error introduced by conducting multiple tests, a single global *F*-test was conducted to test simultaneously whether the impact of any one or more of the outcome measures within a domain is statistically different from zero. Impact models for each of the multiple outcomes within a single domain were estimated simultaneously, and the null hypothesis that the impacts on all of the multiple outcomes are zero was then tested.

Although we adjust for multiple comparisons within a single domain, we do not make adjustments across substantive domains. In other words, we do not combine all outcomes in the study in order to conduct only one single hypothesis test for the study, nor do we combine all three student outcomes (expressive vocabulary, academic knowledge, and listening comprehension) in order to conduct a single test of the K-PAVE impact on students. Schochet (2008b) does not recommend adjusting for multiple comparisons across domains, because doing so produces unnecessarily large reductions in statistical power. Consequently, we maintain the substantive distinction between expressive vocabulary as the primary target of the K-PAVE intervention and the other student outcomes as secondary targets, and we do not adjust for multiple comparisons across the primary and secondary student outcomes.

Because the one primary research question was addressed using a single outcome measure—the EVT–2 standard score—no adjustments for multiple comparisons were required (Table 2.14).

Table 2.14 Primary research question, outcome variable, and need for adjustment for multiple tests

Primary research question	Outcome variable	Adjustment for multiple tests required
1. What is the impact of K-PAVE on students' expressive vocabulary in kindergarten?	• Expressive Vocabulary Test–2 standard score	No

Each of the three secondary research questions addressed in the study represents a separate domain. For some domains, impacts were tested on multiple outcome variables as shown in Table 2.15.

Table 2.15. Secondary research questions, outcome variables, and need for adjustment for multiple tests

Secondary research question	Outcome variables	Adjustment for multiple tests required
2. What is the impact of K-PAVE on students' listening comprehension and academic knowledge in kindergarten?	▪ Listening comprehension standard score ▪ Academic knowledge standard score	Yes
3. What are the impacts of K-PAVE on kindergarten vocabulary and comprehension instruction?	▪ Vocabulary and comprehension support composite ▪ Instructional support ▪ Emotional support	Yes
4. What is the impact of K-PAVE on instructional time use in kindergarten?	Proportion of observation cycles in which language arts content other than vocabulary, comprehension, or oral language is covered	No

For research questions 2 and 3, if the null hypothesis of zero impact of K-PAVE on all of the outcomes is not rejected (for example, no impact on listening comprehension and no impact on academic knowledge), then impacts on individual outcomes within the domain will not be investigated. However, if the null hypothesis of no impact on all outcomes within the domain is rejected, based on the global *F*-test, then the impact on each individual outcome measure will be tested. This within-domain exploratory analysis does not provide confirmatory evidence about the domain as a whole, but it provides information that can help interpret the global findings (Schochet 2008b).

Sensitivity analyses

The following sensitivity analyses were conducted (see Appendix N for details):

- Estimating a baseline model with three levels (school, classroom, student) and no covariates other than the intervention indicator in the level 3 equation and the baseline outcome measure (pretest score) in the level 1 equation.
- Estimating impacts without imputing missing values for outcome variables and pretest scores by single stochastic regression imputation, to test the sensitivity of findings to regression imputation compared with casewise deletion.
- Estimating impacts without imputing missing values for covariates other than pretest scores imputed by dummy variable estimation (see below for an explanation of this method), to test the sensitivity of findings to the dummy variable approach compared with casewise deletion.
- Estimating impacts without imputing student test scores for tests that were incomplete due to administration errors, to test the sensitivity of findings to the imputation of incorrectly administered tests compared with not scoring incomplete tests and allowing values to be missing.

- Estimating impacts without students whose baseline assessments were conducted one to three weeks late,[35] to test the sensitivity of findings to late baseline testing compared with baseline testing conducted prior to K-PAVE training.
- Estimating impacts without outliers, to test the sensitivity of findings to a few influential cases compared with the exclusion of values with large studentized residuals[36] (with an absolute value greater than three).
- Estimating impacts without including students with a raw score of 0 on either a pretest or a posttest, to test the sensitivity of the findings to treating a raw score of 0 as student non-response rather than student inability to answer the test items. For raw scores of 0, it is impossible to know whether the student was unable to answer the test items, in which case a raw score of 0 is valid, or if the student was refusing to complete the test, in which case the score would be treated as missing rather than 0.
- Estimating impacts without weighting schools to adjust for the loss of one school that dropped out of the study.

Chapter 4 reports on cases where the results of the sensitivity analyses deviate from the finding of the main model. (Appendix N contains tables comparing impact estimates from all models.) Sources of sensitivity are reported for impacts on individual outcomes, providing transparency about analytic decisions (missing data imputation) or other factors (delayed baseline testing) that may influence whether an impact is found to be statistically significant.

In conducting global *F*-tests for multiple outcomes within a single domain, sensitivity analysis results were considered when deciding whether to conduct separate tests of impacts on individual outcomes. Specifically, if a global *F*-test was not statistically significant in our main model, but we found that the null hypothesis was rejected in the preponderance of models estimated in the sensitivity analyses, we separately estimated the impact of K-PAVE on the individual outcomes within the domain. In other words, if the results of a global-*F* test were sensitive to different procedures for handling covariate adjustment, imputation of missing data, delayed baseline testing, and student crossovers, we conducted separate tests of impacts on individual outcomes. In this scenario, the Bonferroni correction was applied to the individual significance tests. If the null hypothesis was rejected for the global *F*-test, tests of the individual impacts were conducted without a Bonferroni correction. However, if the *p*-value for the global *F*-test in the main model was above .05 but below .05 in the preponderance of sensitivity models, a Bonferroni correction to the individual hypothesis tests was added to ensure adequate protection against the type I error introduced by conducting separate hypothesis tests for the multiple outcomes.

Methods for handling missing data

The following types of missing data occurred: missing information for teachers and students on demographic covariates, missing test scores for students because they were absent

[35] Late baseline testing occurred only in the control group.
[36] Studentized residuals are calculated by dividing a raw residual (i.e., measured in nominal units of the variable) by the estimated standard deviation for the distribution of raw residuals. Studentized residuals measure the deviation between observed and predicted values in standard deviation units rather than in the nominal units of the variable.

during test administration, incomplete or incorrect administration of the test battery, absence of a classroom observation in the school that dropped from the study, a failure to submit a completed read-aloud observation instrument for one classroom, and a refusal by some teachers to be audio-recorded. Missing data imputation was done separately for intervention and control groups, as described in Appendix M.

For missing covariates other than pretest scores (such as student characteristics and teacher characteristics), a dummy variable adjustment was employed for imputation (Puma, Olsen, Bell, & Price 2009). Missing cases of a variable were all set to a constant value. In addition, the analysis included an indicator variable identifying observations for which the true value of the covariate was missing. The dummy variable adjustment was applied to all missing student, school, and teacher covariates except missing the pretest scores. Table 2.16 shows the percentage of students, teachers, and schools for which at least one covariate was missing.

Table 2.16 Number and percentage of students, teachers, and schools missing covariate data

Missing covariate data	Analytic sample *n* (%)
Missing at least 1 student covariate	21 students (1.6)
Missing at least 1 teacher covariate	4 teachers (3.1)
Missing at least 1 school covariate	7 schools (10.9)

For student assessments both missing pretest scores and missing posttest scores were imputed using single stochastic regression (Puma, Olsen, Bell, & Price 2009). A multiple regression model, adjusted for the multilevel structure of the data, was used to estimate predicted values for each pretest or posttest with missing values. Predictors included all other information collected (including pretest scores, posttest scores, and covariates). For each missing score a randomly selected residual was added to the predicted value from the regression model to obtain an imputed value. The total number of cases missing either a pretest or a posttest score for each student assessment is shown in Table 2.17.

Table 2.17 Number and percentage of students missing either pretest or posttest assessment, by intervention and control group

	Intervention group (*n* = 596)	Control group (*n* = 700)
Assessment variable	*n* (%)	*n* (%)
Expressive vocabulary	36 (6.0)	58 (8.0)
Academic knowledge	47 (7.9)	59 (8.4)
Listening comprehension	50 (8.4)	67 (9.6)

The same single stochastic regression imputation approach was used for a classroom instruction posttest variable—vocabulary and comprehension support—with a single missing value. There were no missing values for any other baseline or posttest classroom instruction variables.

CHAPTER 3: IMPLEMENTATION OF THE INTERVENTION

OVERVIEW

The study addresses two important questions about the implementation of K-PAVE:

1. What was the fidelity of implementation of K-PAVE, as implemented by regular kindergarten teachers as a supplement to their ongoing literacy instruction?
2. What impacts did K-PAVE have on classroom instruction?

The first question is about documenting the extent to which the training and support that was provided for K-PAVE teachers resulted in implementation of the instructional strategies that accurately represented the objectives of the curriculum. The second question is about whether K-PAVE, at the level with which it was actually implemented in the treatment classrooms, resulted in significant and meaningful differences in instruction between treatment and control classrooms.

The study used classroom observations to address both of these questions, but the methodologies in each case were different, in terms of measures, observers, data collection systems, and analysis. Table 3.1 provides an overview of our approach to each question. The findings on fidelity of implementation are presented in Chapter 3. The findings on impacts on classroom instruction are presented in Chapter 4.

DESIGN OF THE K-PAVE PROGRAM

K-PAVE is a kindergarten vocabulary instruction program designed to supplement regular classroom literacy instruction. K-PAVE has a standard protocol and materials for training and supporting teachers. (See Appendix P for a list of K-PAVE materials provided to teachers and Appendix Q for a sample K-PAVE weekly instructional unit.) As described in Chapter 1, K-PAVE was adapted for kindergarten from PAVEd for Success (PAVE), an earlier preschool version. The adaptations involved eliminating instructional components that kindergarten classrooms were presumed to cover and adapting the curriculum materials to focus on vocabulary appropriate for kindergarten, using the Mississippi social studies and science frameworks as a source of vocabulary words. (Appendix R includes a list of the 240 K-PAVE target vocabulary words.) The teacher training and support protocol was also revised. The initial training was shortened, since K-PAVE covered fewer instructional components. PAVE used district literacy specialists trained in the curriculum to provide follow-up support through in-school coaching. Follow-up implementation support in the K-PAVE study did not include such coaches. Instead, members of the implementation team visited each teacher twice to observe and provide feedback and coaching, and telephone conferences were held between the developer and teachers.

Table 3.1 Methodologies for addressing questions on implementation of K-PAVE

	Sample	Measure	Alignment of measures with K-PAVE[a]	Reliability	Observers	Data collection	Type of analysis
Fidelity of implementation	K-PAVE classrooms	K-PAVE Fidelity Observation	Completely aligned	Not assessed	K-PAVE training and implementation support team[b]	At weeks of 17-20 of 24-week K-PAVE program	Descriptive
Classroom instruction impacts	K-PAVE and control classrooms	Vocabulary Record	Highly aligned with K-PAVE Explicit Vocabulary Instruction component	Observers trained to agreement with master coding	Independent observers	At weeks 22-24 of 24-week K-PAVE program	Comparison of means for K-PAVE and control teachers
		Read Aloud Profile	Highly aligned with K-PAVE Interactive Book Reading component	Observers trained to agreement with master coding	Independent observers	At weeks 22-24 of 24-week K-PAVE program	Comparison of means for K-PAVE and control teachers
		Classroom Assessment Scoring System (CLASS K--3): Instructional Support	Aligned with all 3 K-PAVE components	Observers trained to agreement with master coding	Independent observers	At weeks 22-24 of 24-week K-PAVE program	Comparison of means for K-PAVE and control teachers
		Classroom Assessment Scoring System (CLASS K--3): Emotional Support	Aligned with K-PAVE adult conversations component	Observers trained to agreement with master coding	Independent observers	At weeks 22-24 of 24-week K-PAVE program	Comparison of means for K-PAVE and control teachers

a. *Alignment* is defined as the extent to which the constructs being measured are similar to or overlap with the key components of the K-PAVE curriculum. *Completely aligned* indicates that the instrument measures all teaching strategies explicitly identified as part of the K-PAVE curriculum. *Highly aligned* indicates that the instrument measures one or more but not all key teaching strategies explicitly identified as part of the K-PAVE curriculum and also may measure other teaching behaviors that are consistent with but broader than the K-PAVE components and teaching strategies. *Aligned* indicates that the instrument does not measure K-PAVE teaching behaviors but focuses on other teaching behaviors that are consistent with K-PAVE but are more broadly focused.

b. The team included the K-PAVE developer and staff from the University of Georgia and from the Regional Educational Laboratory Southeast.

K-PAVE is designed to be implemented by kindergarten teachers as a supplement to the regular instructional environment. K-PAVE is not currently available from a publisher, but the developer has negotiated a contract with a publisher, and persons interested in the program should contact the developer for more information.

The K-PAVE vocabulary instruction program and teacher training protocol are described briefly below. More detailed descriptions are provided in Appendices P–T.

K-PAVE PROGRAM COMPONENTS AND TEACHING STRATEGIES

K-PAVE is designed to increase students' vocabulary knowledge.[37] The program is built around 240 target words introduced in 24 weekly units of 10 target words per unit. The target vocabulary words are explicitly taught to students and then reinforced through repeated exposure in multiple contexts: storybook reading, extension activities, and classroom conversations and discussion. Storybooks are selected that include the target words; interactive reading techniques are used to engage students in active discussion of the meaning of the words. Small group extension activities use the target words in literacy, math, or other games or exercises to reinforce students' understanding of the meaning of the target words. Finally, teachers are encouraged to use the target words in conversations and discussions with students throughout the day.

K-PAVE is built around three instructional components to increase student knowledge of targeted and nontargeted vocabulary and promote oral language skills: Explicit Vocabulary Instruction (labeled "New Vehicles"), Interactive Book Reading (labeled "CAR Talk"), and Adult-Child Conversations (labeled "Building Bridges"). The K-PAVE teacher training materials specify teaching strategies for delivering each component, including guidance on their intensity and frequency.

Explicit Vocabulary Instruction (New Vehicles)

The 10 target words introduced each week are associated with a common theme. The themes for the current study were selected to align with the state kindergarten thematic units from the Mississippi science and social studies frameworks to respond to state concerns that the vocabulary instruction be consistent with other state education objectives. The target words are part of the general academic knowledge that all students are expected to learn.

As part of K-PAVE, teachers are trained in five strategies for the Explicit Vocabulary Instruction component: quick definitions, novel-name nameless-category (N^3C) strategy, repeated exposure to the words embedded in book reading, extension activities, and teacher-student conversations or discussions using the words. Each is defined below. K-PAVE also uses the family to reinforce vocabulary learning by sending home a list of the week's target vocabulary words for families to use in their conversations with their children.

[37] Information describing K-PAVE is from the *Kindergarten PAVEd for Success Teacher's Guide* (SERVE Center at the University of North Carolina at Greensboro & University of Georgia, June 2008).

Quick definitions. Quick definitions are simple, age-appropriate definitions of the target words. Teachers are trained to provide these definitions when introducing the words using picture cards or when encountering the words in the book reading.

Novel-Name Nameless-Category (N^3C) strategy. The N^3C strategy involves placing a picture or object that is unknown to the students among pictures or objects with which the students are already familiar. For example, a teacher introducing the new target vocabulary word *kiwi* would show students pictures of an apple, an orange, and a kiwi and ask, "Which one is the kiwi?" Because the students are familiar with the other two fruits and know that their names are not *kiwi*, they would point to the correct picture. In this way, students begin to associate the new word with a new object or picture. The repeated use of the N^3C strategy not only helps students learn the meaning of the target words, but it also is hypothesized to motivate students to adopt this strategy in their independent learning of new vocabulary.

Book reading and re-reading to reinforce target vocabulary words. For each of the 24 units, K-PAVE materials include two books, one fiction and one nonfiction, that include the target words. Embedding the new vocabulary in stories provides context to help students learn the meaning of the words. Teachers are trained to read the same selected texts with their students at least twice during the week.

Extension activities to reinforce target vocabulary words. Teachers are trained to use extension activities that integrate each unit's 10 weekly target vocabulary words into learning centers in the classroom. These activities enable students to practice the new words and make them part of their permanent vocabulary. Each unit includes suggestions for extension activities, and teachers are encouraged to develop their own activities that embed vocabulary practice into free play, center activities, and other ongoing classroom procedures. Consider an example for the target word *radish*. Early in the week, the teacher would introduce the word by showing students a picture of a radish and providing a short definition. Following up later with the N^3C strategy, the teacher would ask students to point to the correct picture among pictures of a radish, carrot, and tomato (the last two being words students already know). During the week, the teacher would read the books for that week's unit, which would include the word *radish.* In small group activities, the teacher might use radishes in the math center as manipulatives. In conversations with students during the week, the teacher might talk about having put a radish in her salad. The teacher could also bring in radishes for students to taste.

Interactive Book Reading (CAR Talk)

K-PAVE trains teachers in a method of reading aloud to students—called Interactive Book Reading—that is intended to enhance students' comprehension of the words and concepts in the texts. K-PAVE also provides guidance to teachers on the ideal frequency of reading aloud with students.

Interactive reading. K-PAVE methods of Interactive Book Reading are intended to actively engage students in discussion and promote their understanding of the concepts and vocabulary in the books. In K-PAVE, Interactive Book Reading involves repeated reading of the same book. As the same book is read more than once, the teacher is trained to move from introducing the contents of the book to asking students complex comprehension questions about the book. K-PAVE Interactive Book Reading has three steps:

1. When introducing a new book, the teacher looks through the book with the students without reading the text and may ask students to predict what the book will be about by looking at the pictures.
2. The teacher talks with students about the book during the read-aloud, using three types of questions at least twice for each type (for a total of at least six questions).

- Questions that ask students to label, recall, describe, and locate allow students to demonstrate competence with skills they have already mastered. The goal is to give students a sense of success and confidence.
- Questions that require students to practice abstract thinking by asking students to judge, contrast, compare, summarize, predict, define, take another point of view, or solve problems.
- Questions that build comprehension by relating the book to students' experiences.

3. Teachers are instructed to use more complex questions as the students become more familiar with the material.

Frequency of reading aloud. K-PAVE guidance directs teachers to read the same book at least twice during the week, whether with the whole class or with small groups or individual students. K-PAVE training recommends that each student be read to at least three times a week, either individually or in small groups. A tracking tool enables teachers to record when they read with students, to ensure that they have read at least three times each week to each student.

Adult-Child Conversations (Building Bridges)

K-PAVE trains teachers to engage in extended conversations with individual, pairs, or small groups of students to help them develop strong oral vocabulary skills. Teachers are trained to engage students in extended talk on topics chosen by the students and to be responsive, active listeners, showing interest in and acceptance of students' contributions. Teachers are also trained to use a broad array of new words, to help students increase their word knowledge. As teacher-student conversations accumulate, exposure to more frequently used words may contribute to students' own productive use of such words, offering students extensive opportunities to develop an expressive vocabulary as well as a receptive vocabulary.

Strategies for teacher-student conversations. K-PAVE provides explicit guidance on teacher-student conversations that build oral language skills and vocabulary:

- The student is the conversation leader (teachers show interest and encouragement; time between conversational exchanges is increased to provide more time for the student to speak).
- Teachers model complex language through extension and linguistic recasting of students' statements.
- Teachers reinforce vocabulary development by introducing words that are not part of the students' typical working vocabulary into the conversation.
- Teachers encourage students by using open-ended questions.

- Teachers model good conversational skills and encourage students to use them (not interrupting, showing interest in topics others bring up, including everyone in the conversation).

Frequency of teacher-student conversations. Teachers are encouraged to find opportunities throughout the day for conversations of at least five minutes' duration. Each student is expected to have at least three opportunities each week to engage in conversation with an adult in the class, alone or as part of a small group. K-PAVE offers teachers a tracking tool to ensure that every child meets this goal.

K-PAVE TEACHER TRAINING AND SUPPORT

The K-PAVE model of teacher training and support includes three activities (Table 3.2):[38]

- An initial large group training of intervention teachers and assistant teachers, led by the K-PAVE developer and a team from the University of Georgia.
- Three small-group follow-up telephone conference calls with teachers, led by the developer, to reinforce key elements of classroom instruction and to discuss implementation challenges and strategies for overcoming them.
- Up to three rounds of classroom observation to document whether the key elements of K-PAVE were being implemented as intended, with follow-up remediation discussions with teachers about elements rated as not fully implemented. The classroom observations and remediation support were conducted by an implementation team composed of the developer, University of Georgia trainers, and members of the Regional Educational Laboratory Southeast (who attended the initial K-PAVE training but were not involved in collecting or analyzing data for the impact study). All of the intervention teachers were observed in the first two rounds of classroom observations. The third round was offered only to teachers who had low fidelity scores on the second observation and who had participated in all three of the follow-up phone calls.

[38] To encourage teachers to participate fully in all training and support activities, the study provided continuing education credits for participation in the training activities.

Table 3.2 Schedule of K-PAVE teacher training and support activities

Activity/responsibility	Sample	Dates
Initial group training on K-PAVE/ Developer and staff from the University of Georgia (UGA)	Teachers, aides from all intervention classrooms	• September–October 2008 intervention schools assigned to attend one of three training sessions • After all consents are completed and after baseline student assessments
First fidelity observation using Fidelity Checklist and remediation based on scoring of Fidelity Checklist/ K-PAVE training and implementation support team	All intervention classrooms	September–October 2008 immediately following initial group training on K-PAVE
Follow-up telephone conferences (three calls per teacher)/ Developer and staff from UGA	Teachers, aides from all intervention classrooms invited to attend in small groups (10 or fewer teachers per call)	• October 2008, December 2008, January 2009 • Approximately weeks 4–5, 9–10, and 14–15 of K-PAVE
Second fidelity observation and remediation (see above)	All intervention classrooms	• February–March 2009 • Approximately weeks 17–20 of K-PAVE
Third fidelity observation and remediation (see above)	Intervention classrooms with low fidelity scores in second observation	• March 2009 • Approximately week 21 of K-PAVE

Initial training workshop

The training workshop was designed to last two days for teachers and one day for assistant teachers (see Appendix S for the agenda). The workshop was organized around the three key components of K-PAVE (New Vehicles, CAR Talk, and Building Bridges). The workshop also included time for teachers and assistant teachers to prepare sample units for implementing K-PAVE when they returned to their classrooms; the trainers were available to provide technical assistance. Teachers received a teacher's guide and materials, including 48 children's books, picture cards for the 240 target vocabulary words, and cards for the N^3C strategy. Assistant teachers were invited to attend the training with the classroom teachers. When assistant teachers could not attend the training, teachers decided whether and how to enlist the help of assistant teachers in implementing the K-PAVE instructional elements.

The initial training workshop was scheduled for the beginning of the 2008/09 school year, immediately following random assignment of schools. The workshop was conducted three times, each time for teachers from a different set of the intervention schools (see Table 3.2).[39]

As a supplement to the regular literacy curriculum, K-PAVE is expected to be integrated into ongoing classroom instruction. Teachers have broad latitude to determine the best strategy

[39] Staggering the initial workshop was necessitated by the fact that the consent process for and selection of kindergarten teachers within schools occurred on a rolling basis. Schools were invited to one of the three workshops, depending on the date when the school completed its consent and selection process.

for adding K-PAVE activities to their instructional schedule. For example, the training on the teacher-student conversations in the Building Bridges component indicates only that the conversations should take place whenever the teacher has time to focus on an individual student or small group of students for at least five minutes. Similarly, teachers are encouraged to schedule the large and small group read-alouds at any time. Although the direct instruction on the target vocabulary and the associated extension activities would logically fit into the scheduled literacy block, teachers are left to determine the best times for these K-PAVE activities.

Follow-up telephone conference calls

Three one-hour follow-up conference calls led by the developer provided continuing support for teachers. Scheduled to occur every four to five weeks after the initial training workshop, the conference calls were voluntary. Each teacher was offered a time slot to participate in the three conference calls so that no more than 10 teachers were scheduled for each call. The three phone calls focused on the three major components of K-PAVE: Explicit Vocabulary Instruction (New Vehicles), Interactive Book Reading (CAR Talk), and Adult-Child Conversations (Building Bridges). Each telephone call included discussion about general implementation challenges, implementation issues related to the specific component that was the focus of the call (e.g., Building Bridges), and any additional assistance teachers wanted from the developer. (Appendix T describes the protocol for the conference calls, including specific questions that were asked in the call.) The telephone conference calls were not explicitly tied to the fidelity observations.

Classroom observations and associated remediation

Up to three rounds of coaching on K-PAVE were tied to the results of a 90-minute observation of the classroom by a member of the K-PAVE training and implementation support team.[40] Observers completed a structured Training Fidelity Checklist and Teacher Observation Follow-Up Meeting Protocol (copies in Appendix H).

Immediately following the observation and scoring, teachers were offered a 20–30-minute one-on-one meeting with the observer for follow-up technical assistance or remediation. As stated in the protocol, the purpose of the remediation meeting was to "use timely and elaborative feedback to reinforce good performance, shape and strengthen more tenuous performance, and investigate, explain, and possibly remediate poor or absent behavior" based on items of the Training Fidelity Checklist (Follow-up Meeting Protocol, p. 1). The observers were trained to do the following:

- Provide feedback on any practice that was missing from the teacher's performance or that was performed incorrectly.
- Review with the teacher the relevant section in the Teacher's Guide.

[40] The team included the developer and staff from the University of Georgia and from the Regional Educational Laboratory Southeast.

- Ask the teacher to identify any difficulties encountered or to discuss ideas that were unclear or that presented the greatest challenge.
- Elaborate or provide additional examples to illustrate concepts that appear unclear to the teacher.
- Demonstrate any procedure that the teacher appears to be having a problem with and give the teacher a chance to practice it.

The first observation and remediation session was scheduled to immediately follow the initial group training. The second session was scheduled to follow the last of the three telephone conference calls for a teacher (around 15 weeks into the 24-week intervention period). The third round of fidelity observations was planned only for teachers who, based on the second observation, were still not implementing with fidelity but who had chosen to participate in all three of the conference calls and who wanted a third observation and remediation. This third round was scheduled to occur shortly after the second observation, at approximately week 17.

PROCEDURE FOR ASSESSING FIDELITY OF IMPLEMENTATION OF K-PAVE

Methodology for evaluating fidelity of implementation

The study evaluated the fidelity of implementation separately for the K-PAVE teacher training and support activities and classroom activities.

Fidelity of K-PAVE teacher training and support. Assessment of the fidelity of implementation of the K-PAVE teacher training and support was based on records of attendance and participation of intervention teachers and assistant teachers in the initial training, teacher participation in the three rounds of follow-up telephone conference calls, and teacher participation in the first two rounds of classroom observations and remediation. Table 3.3 summarizes the methods and data sources used to measure fidelity of implementation for training and support activities. Seven training and support activities were open to teachers: (1–2) initial training for teachers and assistant teachers, (3–5) three telephone conference calls, and (6–7) the first two in-class observation and remediation sessions.[41] Each classroom received a fidelity score (ranging from 0-7), indicating the number of training and support activities in which they participated.

[41] Participation in the third round of classroom observations and remediation was voluntary and therefore was not counted as part of the fidelity score.

Table 3.3 Methodology for assessing fidelity of K-PAVE training and support for teachers in intervention schools

Implementation component	Data source	Responsible staff
Initial workshop	Teacher and assistant teacher attendance records	Developer training staff and Regional Educational Laboratory Southeast staff
Follow-up telephone conference calls	Teacher participation records	Developer training staff
First two rounds of classroom observation and remediation sessions	Records of completed remediation sessions (two rounds) based on Training Fidelity Checklist and Teacher Observation Follow-Up Meeting Protocol	Developer training staff and Regional Educational Laboratory Southeast staff

Fidelity of implementation of K-PAVE classroom activities. Assessment of the fidelity of K-PAVE implementation was based on classroom observations using the Training Fidelity Checklist. The checklist was designed by the K-PAVE developer to rate implementation of the 12 key teaching strategies for the three main program components (New Vehicles, CAR Talk, and Building Bridges). Table 3.4 lists these teaching strategies and the criteria for determining fidelity of implementation. Each teaching strategy is rated on a 0/1 scale, where a score of "1" indicates that there was evidence that the strategy was implemented. As shown in Table 3.4, for 5 of the 12 K-PAVE strategies, ratings are based on the presence or absence of the specific instructional behavior. For the other seven strategies, ratings involve an assessment of the quantity or duration of specific activities. Although the fidelity checklist does not include ratings of the quality of implementation of instructional strategies, the checklist does indicate whether a teacher demonstrates at least the minimum level of implementation of key elements of K-PAVE.

The Training Fidelity Checklists were completed based on direct classroom observation. The fidelity observations were scheduled to coincide with the regular literacy instruction block. Intervention teachers provided the observers with a class schedule that detailed when the literacy block of instruction was typically taught. The observations were scheduled at a time that would allow the fidelity observers to evaluate whether teachers were implementing the K-PAVE components. Teachers were occasionally asked to carry out their literacy block at a time other than their typical time to allow observers to maximize the number of observations that could be carried out at a given school. In the second round of observations, because some teachers reported in follow-up telephone conference calls that they were integrating K-PAVE into their science and social studies blocks in the afternoon, fidelity observations were split across both mornings and afternoons. This means that if K-PAVE was being implemented during afternoon instructional activities, the K-PAVE instruction would be captured during the observation. As shown in Table 3.4, for 5 of the 12 items on the Fidelity Checklist, teacher reports rather than direct observation could be used as partial or complete evidence of implementation. Based on the classroom visit, an observer assigned each intervention classroom a score of 0-12, corresponding to the number of teaching strategies rated as being implemented.

Classroom observers were trained to record detailed field notes to describe the evidence on which they based their fidelity ratings, including recording verbatim any instances of teacher use of linguistically complex language, different question types, elaborations of students' speech or vocabulary recasting, and any mention of vocabulary as a topic. (Copies of the coding form and full coding instructions for the Fidelity Checklist are provided in Appendix H.) Subsequent to the observation-based fidelity rating, a member of the developer's team reviewed the field notes from each classroom observation and independently assigned a score for the implementation of each of the 12 teaching strategies. The classroom receives a score of 0–12 based on the number of teaching strategies rated as being implemented. Agreement between the implementation ratings based on the classroom visit and those based on the field notes was calculated using Cohen's kappa statistic. Kappa was .98 across the 12 instructional strategies, and ranged from .92 to 1.00 for the individual strategies. Discrepancies in the fidelity scores on the individual instructional strategies that were assigned by the observer and by the reviewer arose in 1% of the ratings (7 of 720 ratings across 60 intervention teachers). According to the scoring protocol, differences in the fidelity scores assigned by the observer and the reviewer are resolved by the program developer, based on her review of the field notes. The final fidelity score reflected the developer's resolutions.

Table 3.4 Coding K-PAVE classroom fidelity: training fidelity checklist items and coding protocol

K-PAVE component and teaching strategy	Definitions	Criterion for scoring fidelity based on observation[a]	Other Allowable Documentation
Explicit Vocabulary Instruction (New Vehicles)			
1. Quick definitions	Teacher supplied definitions of target words or of any words. This would generally occur early in the unit when words are being introduced.	Score as "1" if any instance of a target or nontarget vocabulary word is observed.	May use teacher self-report of defining target words if it is not the 1st day of a new unit.
2. Sufficient small-group book reading quantity	Small group book readings with seven or fewer students.	Score as "1" if two small group book readings are observed. One small group book reading must be observed.	May use teacher self-report for 1 instance of small group reading.
3. N^3C introduction of vocabulary	• In introducing new words, teacher presents target word picture card in the context of several "known" (nontarget) picture cards. • Teacher should query each by saying "show me [target]" or "who can show me [target]?" or something similar.	Score as "1" if one instance of N^3C instruction is observed.	May use teacher self-report of using $N^3$3 if it is not the beginning of a new unit.
4. Presence of vocabulary targets	• Teacher posts lists, pictures, or props of targets as a group in the room. • Teacher mentions that there are vocabulary words.	Score as "1" if one instance of either of these is observed.	
5. Vocabulary extension activity	• Students sent to centers where they were encouraged to process vocabulary words through some activity.	Score as "1" if one instance of small group extension activities is observed. Notation in lesson plan or child product from the extension activity can be evidence.	
6. At least two extension activities that target vocabulary	• Extension activities target vocabulary.	Score as "1" if at least two groups of children participating in this type of small group activity are observed. One group must be observed.	May use teacher self-report for 1 instance of small group extension activity.

K-PAVE component and teaching strategy	Definitions	Criterion for scoring fidelity based on observation[a]	Other Allowable Documentation
Interactive Book Reading (CAR Talk)			
7. Competence questions	Concrete questions to which students can be expected to know the answer.	Score as "1" if two instances of this type of question are observed.	
8. Abstract questions	Questions asking students to summarize, define, explain, judge, compare, contrast, predict, take another point of view, solve problems, how and why, and questions about students' internal state.	Score as "1" if two instances of this type of question are observed.	
9. Questions related to students' lives	Questions that relate contents of book to students' lives.	Score as "1" if two instances of this type of question are observed.	
Adult-Child Conversations (Building Bridges)			
10. Child-centered conversation	• Conversations should be about students' interests or topics. • Students should be allowed a significant opportunity to talk.	Score as "1" if both characteristics are true for majority of the time in each five-minute conversation observed.	
11. Duration	• Count number of small group or individual conversations of at least five minutes in groups of seven or fewer. • These can be carried out at a variety of times such as following small group academic times, small group book readings or extension activities, or they can be scheduled separately.	Score as "1" if at least two small group conversations that meet criteria are observed. At least one conversation must be observed.	May use teacher self-report for evidence of duration and size of one small group conversation if one conversation is observed.
12. Linguistic complexity	• Complex talk includes complex cognitively complex questions and linguistic recasts. • Vocabulary emphasis includes rare vocabulary and vocabulary recasts.	Score as "1" if at least one instance of both complex talk and vocabulary emphasis is observed.	

a. The fidelity ratings took into account the behavior of both the teacher and the assistant teacher, if there was one present in the classroom at the time of the observation. Observers rotated between the teacher and assistant teacher, focusing for five minutes on the teacher, then for five minutes on the assistant, rotating throughout the observation period.

Source: Training Fidelity Checklist and Handbook for Fidelity Observer.

FINDINGS ON FIDELITY OF K-PAVE IMPLEMENTATION IN THE INTERVENTION CLASSROOMS

Teacher training and support activities

The records of teacher participation in training and support activities show that 40% of teachers attended all training and support activities, and 12% participated in 6 of 7 training and support activities (Table 3.5). However, 27% of classrooms participated in only 1–4 of 7 training and support activities.

Table 3.5 Proportion of intervention teachers participating in K-PAVE training and support

Number of training and support activities attended	Percent of teachers
0	0
1–4	26.7
5	21.7
6	11.7
7	40.0

Note: Sample is teachers in all intervention classrooms (n = 60)
Source: Records from implementation team.

All intervention teachers attended the initial training and received two rounds of fidelity observation, but just 48% participated in all three follow-up conference calls, and 43% missed either two or all three of the calls (Table 3.6). Remediation was indicated for 85% of teachers following both classroom observations, and all teachers who needed remediation received it. At the end of the second fidelity observation, five teachers were identified as implementing fewer than 8 of the 12 instructional strategies with fidelity but also having participated in all three telephone conference calls. The five teachers were offered an additional remediation visit from the developer; two of the teachers accepted the offer.

Table 3.6 Participation of intervention teachers in K-PAVE teacher training and support activities

Activity	Participation (percent)
Initial workshop	
Attendance by lead teacher	100
Attendance by assistant teacher	75
Conference calls	
Participation in three conference calls	48.3
Participation in two conference calls	8.3
Participation in one conference call	21.7
Participation in no conference calls	21.7
Classroom fidelity	
Participation in both sessions	100
Participation in optional third session (five teachers eligible)	40
Remediation	
Received remediation following one observation	8.3
Received remediation following both observations	85.0
Did not need remediation	6.7

Note: Sample is all 60 intervention teachers.
Source: Records from implementation team.

Classroom teaching strategies

Based on the Training Fidelity Checklist completed at the second round of classroom observations (after all training and ongoing support for teachers had been completed), a third of teachers (32%) were rated as demonstrating 10-12 of the 12 K-PAVE teaching strategies (Table 3.7). Another third of the teachers (37%) were rated as demonstrating 8 or 9 of the 12 teaching strategies. There were 6.7% of teachers who implemented four or fewer strategies.

Table 3.7 Number of teaching strategies implemented

Number of instructional strategies implemented	Percent of intervention classrooms
≤ 4 strategies	6.7
5–7 strategies	25.0
8–9 strategies	36.7
10–12 strategies	31.7

Note: Sample is teachers in all intervention classrooms (n = 60).
Source: Training Fidelity Checklists.

Examination of the implementation scores by K-PAVE components indicates that Interactive Book Reading was the only component in which teachers were consistently implementing all of the teaching strategies (Table 3.8). About three-quarters of teachers were

rated as implementing all of the instructional strategies associated with Interactive Book Reading, compared with approximately one-quarter of teachers for the instructional strategies associated with Explicit Vocabulary Instruction and one quarter of teachers for the strategies associated with Adult-Child Conversations. Overall, 13% of teachers implemented all 12 instructional strategies.

Table 3.8 Presence of instructional strategies by K-PAVE program components

K-PAVE component and teaching strategy	Intervention classrooms implementing with fidelity (percent)
Explicit Vocabulary Instruction (New Vehicles)	
Individual instructional strategies	
Quick definitions	96.7
Two small-group book readings	43.3
N^3C vocabulary introduction	50.0
Vocabulary targets posted	93.3
Vocabulary extension activity	86.7
At least two extension activities related to vocabulary	71.7
All 6 teaching strategies	26.7
Interactive Book Reading (CAR Talk)	
Individual instructional strategies	
Competence questions	88.3
Abstract questions	88.3
Relate questions	78.3
All 3 teaching strategies	76.7
Adult-child conversation (Building Bridges)	
Individual instructional strategies	
Child-centered	43.3
Sufficient duration	31.7
Linguistically complex	68.3
All 3 teaching strategies	25.0
All 12 teaching strategies	13.3

Note: Sample is teachers in all intervention classrooms (n = 60). The results shown indicate the percentage of classrooms rated as having the strategy present, as defined by the developer.
Source: Training Fidelity Checklist.

Explicit Vocabulary Instruction. Four of the six teaching strategies within Explicit Vocabulary Instruction were implemented in a majority of classrooms (Table 3.8). The two teaching strategies not implemented in the majority of classrooms were the novel-name nameless-category (N^3C) strategy for introducing vocabulary and twice-daily book readings with small

groups of students. Just over a quarter of classrooms (27%) were rated as implementing all six teaching strategies for Explicit Vocabulary Instruction.

Interactive Book Reading. Each of the three Interactive Book Reading teaching strategies was implemented in a majority of classrooms. Overall, 77% of classrooms were rated as implementing all three teaching strategies.

Adult-Child Conversations. Only one of the three adult-child conversation strategies—"linguistically complex" discussions—was implemented in a majority of intervention classrooms (68%). In more than half of the classrooms, neither of the other two teaching strategies was implemented. Overall, a quarter of classrooms were rated as implementing all three teaching strategies.

SUMMARY OF FINDINGS AND CONCLUDING OBSERVATIONS

All of the elements of the planned K-PAVE teacher training and support were in place, but participation by intervention classroom teachers and assistant teachers varied across the activities. All teachers and 75% of assistant teachers in the intervention group attended the K-PAVE training workshop. All intervention teachers received feedback and remediation twice on K-PAVE implementation, once after group training and again 17 weeks into the 24-week intervention. Participation by teachers was lower for the three follow-up telephone conference calls—43% of the intervention teachers participated in one or no calls.

Based on the Training Fidelity Checklist from the K-PAVE developer, 68% of intervention classroom teachers were rated as implementing at least 8 of the 12 core K-PAVE instructional strategies. However, for two of the broad K-PAVE components (Explicit Vocabulary Instruction and Adult-Child Conversations), fewer than 30% of intervention teachers demonstrated evidence of all the teaching strategies. The variation in the implementation across K-PAVE classrooms, even with the support offered through classroom observations and remediation, is not inconsistent with an effectiveness trial, where the goal is to evaluate impact under typical implementation conditions. In an effectiveness study like this one, teachers are offered all components of the intervention training and support, but the extent to which they implement the program with fidelity is left to the teachers' discretion. The level of support provided to promote high fidelity implementation by all teachers in an efficacy study is not realistic when an intervention is implemented on a larger scale. The goal in an effectiveness study is to test the intervention under conditions similar to those under which a district might adopt and implement the intervention on a larger scale.

One question for the study concerns whether this implementation of K-PAVE is appropriately conceived of as an effectiveness study. Did the version of the curriculum that was tested and the teacher training and support activities represent a "typical" implementation—would implementation of the curriculum be similar if it were taken to scale? This question focuses especially on the teacher training and support activities. The K-PAVE curriculum materials—target vocabulary words, storybooks, and prepared lesson units—are the same as those available to any district that wanted to implement K-PAVE. The training manual is standard and would be provided to districts that purchased K-PAVE. The kind of in-person support provided to teachers in the current study might or might not be implemented by districts adopting K-PAVE.

CHAPTER 4: IMPACT RESULTS

The primary goal of K-PAVE is to improve students' expressive vocabulary in kindergarten. Because the intervention seeks to enhance students' vocabulary not only through Explicit Vocabulary Instruction but also through Interactive Book Reading and teachers' informal conversations with students, K-PAVE is also hypothesized to have a secondary impact on students' academic knowledge and listening comprehension. This chapter reports the results of the analyses of the primary research question on the impact of K-PAVE on kindergarten students' expressive vocabulary, and a secondary research question on the impacts on kindergarten students' academic knowledge and listening comprehension. The rationale for the hypotheses and for the structuring of the research questions as primary and secondary is discussed in Chapter 1 and Chapter 2. Since K-PAVE is hypothesized to achieve these improvements through impacts on teachers' instructional practices (as described in the theory of change presented in Chapter 1), the study also examined impacts on intermediate classroom instruction outcomes.

SUMMARY OF K-PAVE IMPACTS

K-PAVE had a positive and statistically significant impact on the student outcome most directly targeted by the intervention: students' expressive vocabulary at the end of kindergarten. The standardized effect size was 0.14 standard deviation, which corresponds to an additional month of growth in vocabulary during kindergarten (see p. 85 and Appendix W for a discussion of how impacts on students can be translated into differences in age-equivalent scores).

K-PAVE also had a positive and statistically significant impact on one of the two secondary student outcomes: students' academic knowledge at the end of kindergarten. The standardized effect size was 0.14 standard deviation, which corresponds to an additional month of growth in academic knowledge during kindergarten. K-PAVE did not have a statistically significant impact on the other secondary student outcome: students' listening comprehension at the end of kindergarten.

The intervention also had a positive and statistically significant impact on the classroom instruction outcome that was most closely aligned to K-PAVE—vocabulary and comprehension support during reading aloud and other classroom activities. The standardized effect size of the impact on this outcome was 0.82. There was no impact of K-PAVE on two other classroom instruction measures, which were less highly aligned with K-PAVE and represent more general quality constructs—instructional support and emotional support from the CLASS. The fact that K-PAVE resulted in significant differences in instructional practices that support vocabulary and comprehension provides a rationale for why K-PAVE resulted in impacts on students. The theory of action suggests that classroom instructional practices are important mediators through which K-PAVE affects student outcomes, and K-PAVE was found to have an impact on an outcome measuring classroom practices that are hypothesized to be closely linked with fostering students' vocabulary development.

RESEARCH QUESTIONS

Specifically, this chapter presents the results of analyses that address four research questions, one primary:

1. What is the impact of K-PAVE on students' expressive vocabulary in kindergarten?

In addition, there are three secondary research questions:

2. What are the impacts of K-PAVE on students' listening comprehension and academic knowledge in kindergarten?
3. What are the impacts of K-PAVE on kindergarten instructional practices, specifically vocabulary and comprehension support during book reading, instructional support, and emotional support?
4. Does implementation of K-PAVE vocabulary teaching come at the expense of time for instruction in other areas of early literacy (such as concepts of print, phonological awareness, alphabetic knowledge, reading, writing, fluency, and spelling)? Specifically, do K-PAVE teachers spend less time on these nonvocabulary literacy teaching practices compared with control group teachers?

IMPACTS ON KINDERGARTEN STUDENTS' EXPRESSIVE VOCABULARY

This section presents the results of the analysis addressing the primary research question: What is the impact of K-PAVE on students' expressive vocabulary in kindergarten?

The Expressive Vocabulary Test–2 (EVT–2) is a one-on-one, normed, and standardized test for measuring students' expressive vocabulary. Students provide a single-word response to describe a pictured stimulus. Scores were normed using a national sample of children of the same age and were standardized in the norming sample to have a mean of 100 and a standard deviation of 15 points. (Sample means and standard deviations for all outcome measures are presented in Appendix U.)

The estimated impact of K-PAVE on kindergarten students' expressive vocabulary is 1.60 points, which is a standardized effect size of 0.14 standard deviation (Table 4.1).[42] The 95% confidence interval around the impact estimate is 0.43–2.77 points.[43] The estimated magnitude and standard error of the impact remain consistent in all sensitivity analyses (see Appendix N).

[42] Model assumptions of residual normality and homoscedasticity were met for this model and all other models presented in this chapter. Information on the evaluation of model assumptions and examination of influence statistics for all models is in Appendix V. The standardized effect size is calculated by dividing the parameter estimate (measured in the units of the EVT standard score scale) for the intervention indicator by the standard deviation of the control group posttest score.

[43] Although the a priori statistical power analysis suggested that the study would have 80% power to detect an impact on students of 0.26–0.28 standard deviation or higher, the actual minimum detectable effect size for the EVT–2 was 0.14. The study's actual power for detecting impacts on students was higher than anticipated, primarily because the school-level R^2 was higher than assumed in the initial statistical power analysis (see Appendix B).

Table 4.1 Estimated regression-adjusted impact of K-PAVE on kindergarten students' expressive vocabulary

Focus of research question	Regression-adjusted posttest means: Intervention group	Regression-adjusted posttest means: Control group	Estimated impact (standard error)	*p*-value	95% Confidence interval	Effect size[a]
Expressive Vocabulary Test–2 (EVT–2) standard score	93.22	91.62	1.60** (0.59)	.006	0.43–2.77	0.141

** $p < .01$.

Note: The intervention group includes 596 students in 60 classrooms in 30 schools; the control group includes 700 students in 68 classrooms in 34 schools. A three-level model was used to estimate impact, controlling for school-level and student-level covariates.

a. The effect size was calculated by dividing the estimated impact by the standard deviation (11.35) of the control group posttest EVT–2 score.

Students' posttest EVT–2 scores were translated into age-equivalent scores based on the norming sample (see Williams 2007). At posttest the regression-adjusted mean EVT–2 score in the intervention group was equivalent to the average score for children age 5 years and 7 months; the regression-adjusted posttest EVT–2 score in the control group was equivalent to the average score for children age 5 years and 6 months. (See Appendix W for a discussion of how the age-equivalent scores were determined.) Thus, students who received the K-PAVE intervention were one month ahead of students in the control group in vocabulary development.

Hill, Bloom, Black, & Lipsey (2008) suggest examining the magnitude of an impact in comparison to the size of performance gaps. In this study, the average baseline EVT score among students in the study is 0.60 standard deviation below the norm for students their age.[44] By the end of the intervention, students in the control group, in the absence of K-PAVE, are still .56 standard deviations below the norm, while students in the K-PAVE intervention group are .45 standard deviation below the norm. Although students in the intervention group are still nearly half a standard deviation below the norm, they are 20% closer to the norm than students in the control group. In other words, the performance gap has been reduced by 20%.

IMPACTS ON KINDERGARTEN STUDENTS' ACADEMIC KNOWLEDGE AND LISTENING COMPREHENSION

This section presents results of the analysis addressing the secondary research question:

- What are the impacts of K–PAVE on students' listening comprehension and academic knowledge in kindergarten?

The Kaufman Test of Educational Achievement, Second Edition (KTEA–II) Listening Comprehension subtest tests students' ability to answer comprehension questions about short passages read aloud to them. Scores were normed using a national sample of students of the

[44] The average baseline EVT score is 91 points on a normed scale with a mean of 100 and a standard deviation of 15. On average, students score 9 points below the norm, which is 0.60 standard deviations (9 points /15 points) below the norm.

same age and standardized to have a mean of 100 and a standard deviation of 15 points. Individual student scores in the sample range from 56 points to 123 points, with a mean of 89.1 points and a standard deviation of 13.0 points.

The Woodcock-Johnson III/Normative Update (WJ–III/NU) Academic Knowledge subtest measures students' background knowledge in science, social studies, and humanities. An IRT-scale score (*W*-score) was calculated based on a nationally representative norm group. Student scores in the sample range from 396 points to 559 points, with a mean of 455.7 points and a standard deviation of 13.0 points. The *W*-scores correspond to scores on a standard score scale (with normed mean of 100 and standard deviation of 15) that range from 43 points to 131 points, with a mean of 91.7 points and a standard deviation of 11.5 points.

We conducted a global *F*-test to test the null hypothesis that K-PAVE's impacts on kindergarten students' academic knowledge and listening comprehension are both zero. We did not reject the null hypothesis ($F_{2, 96} = 3.08$, $p = .051$). Sensitivity analyses (described in Chapter 2 and Appendix N) revealed that this finding is highly sensitive to four of six issues examined. The four issues involved the handling of: incorrectly administered baseline tests (1 model); late baseline testing (2 models); student crossovers and transfers (3 models); and posttest raw scores of 0 (1 model). Findings were not sensitive to covariate adjustment (1 model) or the handling of missing test scores and covariates (2 models). In 7 of the 10 sensitivity models estimated, we *did* reject the null hypothesis of zero impact on both outcomes, with *p*-values below .050. Thus, the results of the majority of the sensitivity models were not consistent with the result from the main model.

We could not reject the null hypothesis of zero impact on both academic knowledge and listening comprehension in the main model; however, it was rejected in most of the sensitivity models. Therefore, we examined the impact of K-PAVE on each of the two outcomes separately. However, because the *p*-value for the global *F*-test in the main model was above .05, a Bonferroni correction was added to the hypothesis tests to protect against the type I error introduced by conducting separate hypothesis tests for the two outcomes. For each of the two hypothesis tests, a criterion of $p < .025$ was used to identify an impact as statistically significant.

Table 4.2 shows the estimated impact of K-PAVE on listening comprehension and academic knowledge in kindergarten. The estimated impact on listening comprehension of 1.4 points, which is an effect size of .11 standard deviation, is not statistically significant ($t = 1.60$; $p = .11$). The 95% confidence interval, which ranges from -0.35 point to 3.17 points, includes zero. This finding is consistent in all sensitivity analyses (see Appendix N).

There is a statistically significant impact of K-PAVE on academic knowledge ($t = 2.29$; $p = .022$[45]). The impact estimate is 1.95 points, with a 95% confidence interval of 0.25–3.65 points. The standardized effect size of 0.14 indicates that students in intervention schools score 0.14 standard deviation higher on the WJ–III/NU academic knowledge test than do students in control schools. The magnitude and standard error of the estimate remain consistent in the sensitivity analyses.[46]

---

[45] The Bonferroni adjustment for multiple comparisons sets the threshold of statistical significance at .025 when two outcomes are examined within the same domain. A *p*-value of .022 meets this threshold.

[46] In 4 of the 12 sensitivity models testing the impact of K-PAVE on academic knowledge, the *p*-value is above the .025 threshold for statistical significance set by the Bonferroni adjustment. The *p*-values range from .026 to .043, indicating some sensitivity to covariate adjustment, imputation of missing covariate data (two models), and

Students' posttest academic knowledge scores were translated into age-equivalent scores, which were developed on the norming sample (McGrew, Schrank, & Woodcock 2007). At posttest, students in intervention schools had an average age-equivalent score of 5 years and 10 months, and students in control schools had an average age-equivalent score of 5 years and 9 months. (See Appendix W on how the age-equivalent scores were determined.) Thus, kindergarten students who received the K-PAVE intervention were one month ahead of students in control schools in academic knowledge.

Table 4.2 Estimated regression-adjusted impact of K-PAVE on kindergarten students' listening comprehension and academic knowledge

	Regression-adjusted posttest means					
Focus of research question	**Intervention group mean**	**Control group mean**	**Estimated impact[a] (standard error)**	***p*-value**	**95% confidence interval**	**Effect size[b]**
Listening comprehension standard score	90.13	88.72	1.41 (0.88)	.11	–0.35–3.17	0.109
Academic knowledge IRT-based *W*-score	456.48	454.53	1.95* (0.85)	.022	0.25–3.65	0.144

* $p < .025$ (p-value thresholds reflect the Bonferroni adjustment for testing two outcomes in a domain).
Note: The intervention group includes 596 students in 60 classrooms in 30 schools; the control group includes 700 students in 68 classrooms in 34 schools.
a. A three-level model was used to estimate impacts, controlling for school-level and student-level covariates.
b. Calculated by dividing the estimated impact by the standard deviation of the control group posttest (13.0 for listening comprehension and 13.48 for academic knowledge).

IMPACTS ON CLASSROOM INSTRUCTION

- This section addresses the secondary research question about the impacts of K-PAVE on kindergarten classroom instruction, which is hypothesized to improve students' vocabulary and vocabulary-related outcomes: What are the impacts of K-PAVE on kindergarten instructional practices, specifically vocabulary and comprehension support during book reading, instructional support, and emotional support?

The question was addressed in two steps. First, the null hypothesis of no impact of K-PAVE on vocabulary and comprehension support, instructional support, or emotional support was tested. The global F-test found a statistically significant impact of K-PAVE on one or more of these measures of classroom instruction ($F_{3,132} = 5.26$, $p = .002$). This finding remained consistent in all sensitivity analyses (see Appendix N). Second, because the null hypothesis of zero impacts on all three outcomes was rejected, the impact on each type of instructional practice was tested individually.

imputation of scores for incorrectly administered tests; however, for all models, the magnitude of impact estimates (effect sizes ranging from .12 to .16) and standard errors are similar to those in the main model (see Appendix N).

Table 4.3 shows the estimated impact of K-PAVE on the three classroom instruction outcomes in kindergarten.[47] K-PAVE has a statistically significant impact on classroom instruction focused on vocabulary and comprehension support ($t = 3.50$, $p = .0009$). The impact estimate is 0.73 point, with a 95% confidence interval of 0.31–1.14 points. The standardized effect size is 0.82, indicating that, on average, classrooms in intervention schools provided students with vocabulary and comprehension support that was .82 standard deviations higher than classrooms in control schools.[48]

Table 4.3 Estimated regression-adjusted impacts of K-PAVE on classroom instruction in kindergarten

Focus of research question	Regression-adjusted posttest means: Intervention group	Regression-adjusted posttest means: Control group	Estimated impact [a] (standard error)	*p*-value	95% confidence interval	Effect size[b]
Vocabulary and comprehension support	0.498	–0.228	0.726*** (0.21)	.0009	0.31–1.14	0.823
Instructional support	3.69	3.09	0.60~ (0.36)	.097	–0.11–1.31	0.470
Emotional support	5.47	5.31	0.156 (0.20)	.439	–0.24–0.56	0.196

*** $p < .001$ ~ $p < .10$

Note: The intervention group includes 60 classrooms in 30 schools; the control group includes 68 classrooms in 34 schools.

a. A two-level model was used to estimate impact, controlling for both school and teacher covariates at the school level.

b. Calculated by dividing the estimated impact (the raw parameter estimate for the intervention indicator) by the standard deviation of the control group posttest score. The control group standard deviation was 0.882 for vocabulary and comprehension support, 1.28 for instructional support, and 0.80 for emotional support.

As described in Chapter 2, the composite measure of vocabulary and comprehension support was created from four variables: number of words introduced during a book read-aloud, average number of words introduced during other instructional time, number of comprehension supports (such as background information, connections to children's experience, and clarifications) provided during the read-aloud, and number of higher order questions asked during the read-aloud. This composite measure was created for this study, so there is no basis for comparing schools' scores on the measure other than for schools within this sample. In addition, the measure cannot be interpreted in the original units because the four variables were converted

[47] The vocabulary and comprehension support composite has a *z*-score scale, which was standardized on the study sample to have a mean of 0 and a standard deviation of 1. A positive score indicates more vocabulary and comprehension support than the average for the sample, and a negative score indicates less. The instructional support and emotional support variables are measured on a seven-point Likert scale.

[48] The estimated magnitude and standard error of this intervention effect remained consistent across most of the sensitivity analyses (see Appendix N). In all models except one, the standardized effect size was 0.79–0.82; the exception (a model with just two covariates: intervention status and the baseline vocabulary and comprehension support score) suggested a somewhat larger standardized effect size of 0.97. This difference suggests that not controlling for covariates runs the risk of attributing variation in classroom instruction associated with teacher and school characteristics to the K-PAVE intervention.

to a scale with a mean of zero and standard deviation of one in order to weight them equally when summed.

The nominal units of the four component measures of the composite vocabulary and comprehension support can provide some context for interpreting the 0.82 standard deviation difference in classroom vocabulary and comprehension support between intervention and control classrooms (Table 4.4). For example, control group teachers provided comprehension support (provided background information related to the book, made connections to students' experiences, clarified meaning, or asked students to clarify the text) an average of 17 times during a book read-aloud. With a control group standard deviation of 14.5 instances of comprehension support, 0.82 standard deviation would imply 12 instances of providing comprehension support during a read-aloud. If this were the only component of vocabulary and comprehension support measured, an impact of 0.82 could be interpreted to mean that teachers in intervention schools were providing comprehension support to students during book reading 12 times more than teachers in control schools.

Table 4.4 Sample control group means and standard deviations for four components of vocabulary and comprehension support composite measure (*n* = 128 classrooms)

Component	Mean	Standard deviation
Comprehension support during read-aloud (frequency)	17 times	14.7 times
Higher order questions during read-aloud (frequency)	3 questions	3.3 questions
Number of word meanings introduced during read-aloud	4 words	4.2 words
Number of word meanings introduced during other times	1.8 words	1.4 words

For higher order questions during book reading (asking students to analyze, explain, predict, imagine, make inferences, or generate hypotheses), the sample mean for the control group was 3 questions, with a standard deviation of 3.3 questions. If higher order questions were the only component measured, an impact of 0.82 could be interpreted to mean that teachers in intervention schools were asking almost three more higher order questions during book reading than were teachers in control schools.

For the introduction of word meanings—during both book reading and other instructional periods—the sample mean for the group was 4 word meanings introduced during book readings, with a standard deviation of 4.2 word meanings and nearly 2 word meanings at other times, with a standard deviation of 1.4 word meanings. A difference of 0.82 standard deviation translates into three (3.4 words) additional words during book reading and an additional word (1.2 words) during other times of the day.

K-PAVE had no statistically significant impact on the level of instructional support or emotional support provided in the classroom (see Table 4.3). For emotional support, findings remained consistent in the sensitivity analyses (see Appendix N). For instructional support, the findings were sensitive to covariate adjustment and to the imputation of missing values for teacher and school characteristics.[49] Although the null hypothesis of zero impact on instructional

[49] In a model with no covariates other than intervention status and baseline instructional support, the effect size of 0.59 standard deviation is statistically significant ($t = 2.46$, $p = .02$). As noted previously, the difference in magnitude of the impact estimate suggests that without the covariates, the variation in instructional support

support is not rejected in the main model, models without covariate adjustment or without imputation of missing values for covariates offer some indication that K-PAVE may affect instructional support.

In both intervention and control classrooms, the level of instructional support can be considered of moderate quality according to the Classroom Assessment Scoring System (CLASS), with regression-adjusted mean ratings of 3.8 and 3.2.[50] The level of emotional support can be considered to be on the high end of moderate quality, with regression-adjusted mean ratings of 5.5 in intervention classrooms and 5.3 in control classrooms.

Impacts of K-PAVE on instructional time use in kindergarten

- The study also tested whether K-PAVE had an unintended impact on instruction in other areas of early literacy: Does implementation of K-PAVE vocabulary teaching come at the expense of time for instruction in other areas of early literacy (such as concepts of print, phonological awareness, alphabetic knowledge, reading, writing, fluency, and spelling)? Specifically, do K-PAVE teachers spend less time on these nonvocabulary literacy teaching practices compared with control teachers?

The study measured the proportion of 20-minute CLASS observation cycles that included literacy instruction in areas other than vocabulary and comprehension support. Time spent on areas of literacy instruction other than vocabulary and comprehension support was an estimated 1.1 percentage points higher in control classrooms than in intervention classrooms, which is not a statistically significant difference (t = -0.14, p = .89). Thus, the null hypothesis that the impact of K-PAVE on instructional time spent on nonvocabulary literacy instruction is zero cannot be rejected. The magnitude and standard error of the impact estimate remained consistent in sensitivity analyses (see Appendix N).

---

associated with teacher and school characteristics would be attributed to the K-PAVE intervention. In a model estimated without imputing missing values for teacher and school characteristics, the impact was statistically significant (t = 2.14, p =.04), with an effect size of 0.55 standard deviation.

[50] Ratings in the 1–2 range are considered poor quality, in the 3–5 range moderate quality, and in the 6–7 range high quality (Pianta, LaParo, & Hamre 2008).

Table 4.5 Estimated regression-adjusted impact of K-PAVE on amount of literacy instruction in areas other than vocabulary and comprehension in kindergarten

Focus of research question	Regression-adjusted posttest means: Intervention group	Regression-adjusted posttest means: Control group	Estimated impact (standard error)[a]	*p*-value	95% confidence interval	Effect size[b]
Proportion of instructional cycles focused on areas of literacy other than vocabulary or comprehension	0.502	0.513	–0.011 (0.07)	.89	–0.16–0.14	–0.03

Note: The intervention group includes 60 classrooms in 30 schools; the control group includes 68 classrooms in 34 schools.

a. A two-level model was used to estimate impact, controlling for both school and teacher covariates at the school level.

b. Calculated by dividing the estimated impact by the standard deviation of the control group posttest measure (standard deviation = 0.329).

CHAPTER 5: SUMMARY OF FINDINGS AND STUDY LIMITATIONS

This chapter summarizes the findings on the impact of K-PAVE on kindergarten students' expressive vocabulary and related literacy skills and on classroom instruction in ways hypothesized to improve students' vocabulary outcomes. The chapter also describes the main design parameters for the study and the study's strengths and limitations.

EFFECT OF K-PAVE ON EXPRESSIVE VOCABULARY

The study found that kindergarten students in schools using K-PAVE as a supplement to the regular literacy instruction performed better than kindergarten students in control schools on the Expressive Vocabulary Test–2 administered at the end of the school year.[51] The estimated impact on expressive vocabulary is 1.60 points, which represents a standardized effect size of 0.14 standard deviation. The impact is statistically significant, and the 95% confidence interval around the impact estimate is 0.4–2.8 points. An impact of this magnitude means that students who received the K-PAVE intervention are one month ahead in vocabulary development at the end of kindergarten compared with students in the control group.

EFFECT OF K-PAVE ON OTHER VOCABULARY-RELATED LITERACY SKILLS

The study also tested the impact of K-PAVE on two secondary student outcomes related to vocabulary knowledge: academic knowledge (measured with the Woodcock-Johnson III/Normative Update) and listening comprehension (measured with the Kaufman Test of Educational Achievement II). The study found that kindergarten students in K-PAVE schools performed better than students in control schools on the measure of academic knowledge administered at the end of the year.[52] The estimated impact on expressive vocabulary is 1.95 points, which represents a standardized effect size of 0.14 standard deviation. The impact is statistically significant, and the 95% confidence interval around the impact estimate is 0.2–3.7 points. An impact of this magnitude means that students who received the K-PAVE intervention are one month ahead in academic knowledge development at the end of kindergarten compared with students in the control group. The study did not find a statistically significant difference between kindergarten students in K-PAVE schools and students in control schools on the measure of listening comprehension administered at the end of the school year.

EFFECT OF K-PAVE ON CLASSROOM INSTRUCTION

The theory of action for K-PAVE posits that effects on students will be mediated through effects on their teachers' instructional practices. K-PAVE caused a positive and statistically significant impact on one of the three kindergarten classroom instructional practices examined: vocabulary and comprehension support, which includes introducing vocabulary words during

[51] The study conclusions are not sensitive to the inclusion or exclusion of student and school covariates, alternative approaches for imputing missing data, or to the handling of extreme values (outliers).

[52] The study conclusions are not sensitive to the inclusion or exclusion of student and school covariates, alternative approaches for imputing missing data, or to the handling of extreme values (outliers).

read-alouds, introducing vocabulary words throughout the school day, asking higher order questions during read-alouds, and providing comprehension support during book read-alouds. The estimated impact on vocabulary and comprehension support was 0.82 standard deviation. Although the relative contribution to this impact of each of the four components of the measure is unknown, if all contributed equally, the size of this impact is equivalent to introducing three more words during book reading, introducing one more word during other times of the day, asking three more higher order questions during book read-alouds, and providing comprehension support during book read-alouds 12 more times than in control classrooms.

The study tested the impact of K-PAVE on two other classroom instruction outcomes through which K-PAVE is hypothesized to affect students—instructional support and emotional support, as measured by the Classroom Assessment Scoring System (CLASS). K-PAVE classrooms were not significantly different than control classrooms on either outcome measured at the end of the year.

Finally, K-PAVE did not have an unintended effect of reducing the amount of time spent on areas of literacy instruction other than vocabulary and comprehension (such a phonological awareness, alphabet knowledge, print concepts, and decoding) Thus, the positive impact of K-PAVE on vocabulary and comprehension support did not come at the expense of time spent on other literacy instruction.

STUDY PARAMETERS

These results come from the first randomized controlled trial testing the effectiveness of K-PAVE in enhancing expressive vocabulary of kindergarten students. The study employed a cluster randomized design, with approximately 1,300 students nested in 64 schools in 35 districts. The design has strong internal validity, based on the randomization procedure and the low levels of attrition from the sample at all levels over the outcome period. The study is well powered to detect impacts on students, and the multilevel modeling used in the analysis appropriately accounts for the nesting of students in classrooms and schools. The study generated statistically unbiased estimates of the impact of K-PAVE under typical implementation conditions for the Mississippi Delta region for schools that volunteer for this type of study.

The study evaluated the implementation of K-PAVE to assess whether the intervention delivered to students represented a strong version of the program as designed. Results on fidelity of implementation indicated that the teacher training and support activities were implemented as planned and that, as expected in an effectiveness trial, there was substantial variation in implementation across K-PAVE classrooms. A total of 68% of teachers implemented at least 8 of the 12 instructional strategies; 25% of teachers implemented 5–7 strategies, and 7% implemented 1–4 strategies.

The study tests K-PAVE as implemented under typical rather than optimal conditions. K-PAVE was implemented by regular kindergarten teachers as a supplement to their ongoing literacy instruction. This suggests that schools that want to implement K-PAVE could replicate the study's implementation conditions. As indicated in Chapter 1, the developer has negotiated a contract with a publisher to offer K-PAVE and PAVE, which would be called PreK-PAVE. Persons interested in the program should contact the developer for more information. The impact results represent effects when the teacher training and support model includes the elements

described previously – group training; curriculum materials and sample lesson plans; follow-up training; and in-class observation and support. There is no way to know whether this full menu of training and support represents what districts would "typically" implement.

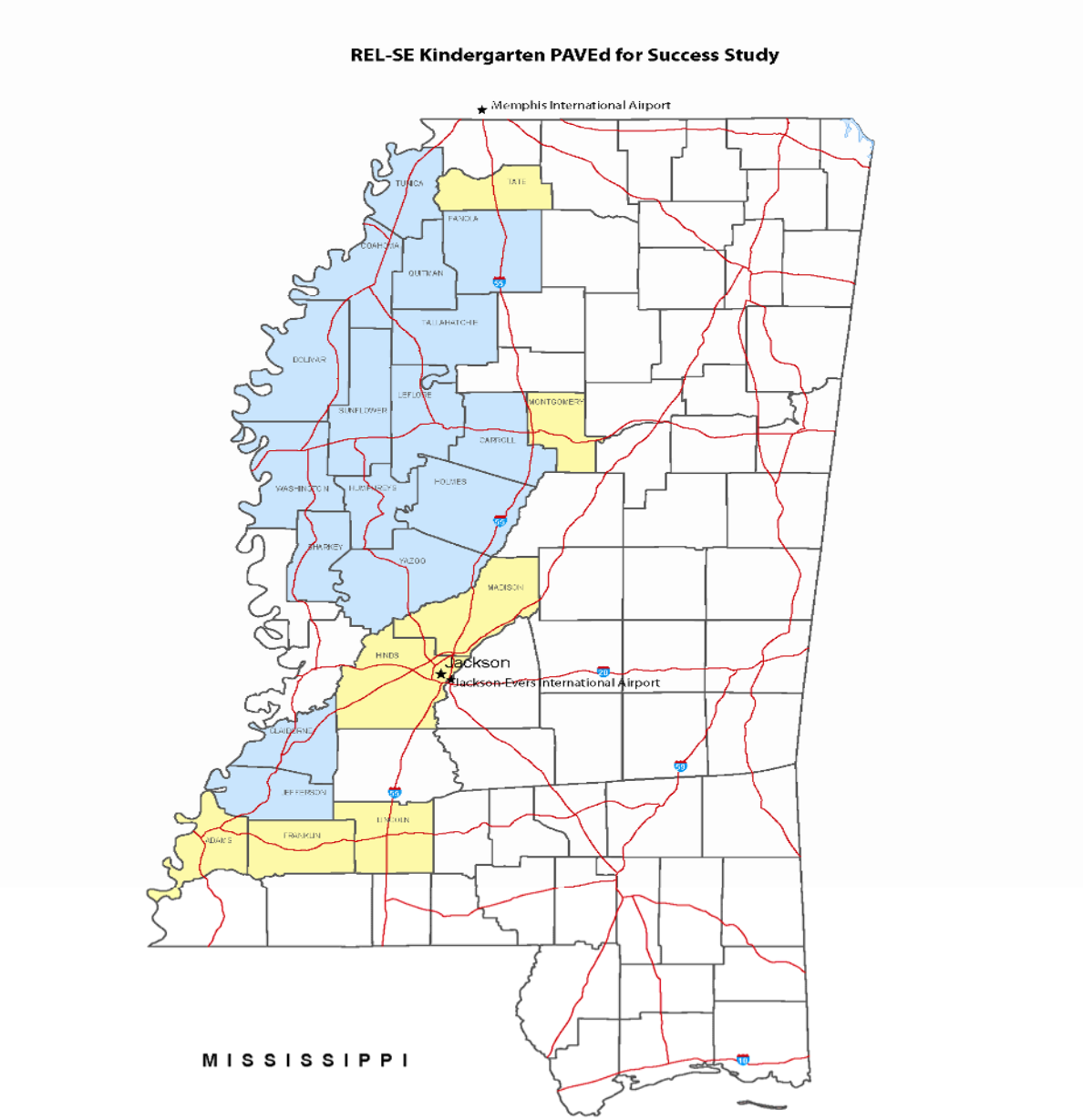
STUDY LIMITATIONS

Although the study is a rigorous test of K-PAVE, with strong internal validity, there are limitations to the generalizability of the findings. First, the results apply to K-PAVE implemented at the kindergarten level. Second, because the sample districts, schools, and students were all in the Mississippi Delta and surrounding counties, findings are not generalizable to other areas. Moreover, the study was not a random sample of eligible schools in the Delta and surrounding counties but of schools that volunteered to participate. However, schools that volunteered were similar to the pool of all eligible schools on a set of measured characteristics (including region, school performance classification, extent of meeting annual expectations for growth in achievement, and eligibility for free or reduced-price meals). Compared with schools that declined to participate, volunteer schools tended to have a higher percentage of African American students, were more likely to be in small towns, and less likely to be located in rural areas.

At this point, the study does not provide evidence on the important question of whether the effects of K-PAVE in kindergarten provide a sustained advantage for the students in intervention schools. Does a significant boost in vocabulary in kindergarten have long-lasting effects on students' vocabulary and related literacy skills in first grade and the formal reading instruction that typically begins in that grade? The evidence of the positive impact of K-PAVE on students' expressive vocabulary and academic knowledge at the end of kindergarten allows the study to address this question, since the study design stipulates that the kindergarten cohort will be followed into first grade if there is evidence of a positive impact of K-PAVE on these outcomes at the end of kindergarten. This question will be examined in a follow-up study.

APPENDIX A. MISSISSIPPI COUNTIES WITH STUDY SCHOOLS, BY COUNTY

Map A1. Mississippi counties with study schools, by county



APPENDIX B. STATISTICAL POWER ANALYSIS

This appendix presents results from a priori statistical power analyses conducted to determine the sample size targets for recruitment.

STATISTICAL POWER FOR DETECTING IMPACTS ON STUDENTS

The a priori statistical power analysis suggested that the study would have 80% power to detect impacts on students' expressive vocabulary of 0.26–0.28 standard deviation or higher. In addition to the details of the statistical power analysis conducted in designing the study, the study's actual statistical power, which was determined based on knowledge of the impact results, is reported.

The study has a cluster randomized design in which schools were randomly assigned to the K-PAVE intervention or a control condition. Two kindergarten classrooms were randomly selected from each school, and 10 students were randomly selected from each classroom. The hierarchical structure of the data led to the use of the following equation to estimate minimum detectable effect sizes (Schochet 2008a):

(B1)

$$MDES(\hat{\beta}_1\ treatment) = Factor(\alpha, \beta, df) * \frac{\sqrt{\frac{\sigma^2_{school}(1-R^2_{school})}{sp(1-p)} + \frac{\sigma^2_{class}(1-R^2_{class})}{(sp(1-p))c} + \frac{\sigma^2_{child}(1-R^2_{child})}{(sp(1-p))cn}}}{\sigma}$$

where $MDES(\hat{\beta}_1\ TREATMENT)$ is the estimated minimum detectable effect size for the treatment impact; $Factor(\alpha, \beta, df)$ is a constant that is a function of the significance level (α), statistical power (β), and the number of degrees of freedom (*df*); σ^2_{school} is the school-level variance in the outcome; σ^2_{class} is the classroom-level variance in the outcome; σ^2_{child} is the student-level variance in the outcome; R^2_{school} is the amount of school-level variance in the outcome explained by covariates; R^2_{class} is the amount of classroom-level variance in the outcome explained by covariates; R^2_{child} is the amount of student-level variance in the outcome explained by covariates; *s* is the number of schools; *p* is the proportion of schools assigned to the treatment condition; *c* is the number of classrooms sampled from each school; *n* is the average number of students sampled from each classroom; and σ is the standard deviation of the outcome measure for the control group.

Based on previous quasi-experimental evaluations of the PAVEd for Success program, which found standardized effect sizes ranging from 0.20 to 0.43, the goal was to recruit a sample that would enable effect sizes at the lower end of this range to be detected. The following assumptions were thus in the statistical power analysis:

- The proportion of total variation in student outcomes at the school level is between .10 and .15 because of the project's focus on recruiting schools exclusively within high-poverty communities (see Hedges and Hedberg 2007).[53]
- The proportion of total variation in student outcomes at the classroom level is .05. It was assumed that variation in student outcomes between classrooms would be relatively low given that kindergarten classrooms are generally not grouped based on ability.
- The pretest explains 50% of the variation between schools in posttest scores (that is, school-level $R^2 = .50$).
- The correlation between any other covariates and the outcome measure, conditional on the value of the pretest measure, is .00 (so these covariates do not influence the school-level R^2).[54]
- The classroom-level R^2 and student-level R^2 were both 0, in order to be conservative in the power calculation.
- By the end of the study, the student attrition rate will equal approximately 20%.
- A two-tailed test of significance will be conducted at the .05 level.
- The desired power to detect effects is .80.

Based on the statistical power analysis, the recruitment target was 60–70 schools, with two classrooms per school[55] and 10 students per classroom.[56] Table B1 indicates the minimum detectable effect sizes for student outcomes anticipated for 60–70 schools, with an intraclass correlation of .10–.15 and based on the above assumptions. The estimated minimum detectable effect size ranges from 0.25 to 0.29.

Table B1. Power analysis summary: minimum detectable effect sizes for student outcomes, by number of schools

School-level intraclass correlation	70 schools	65 schools	60 schools
.10	0.25	0.26	0.27
.15	0.27	0.28	0.29

[53] Based on a compilation of interclass correlations in group randomized evaluations of student achievement, Hedges and Hedberg (2007) find that the intraclass correlation for low–socioeconomic status schools is approximately .19 and for low-achievement schools is approximately .09.

[54] A higher school-level R^2 (that is, additional explanatory power for school-level covariates other than for pretest) would reduce minimum detectible effect sizes. It is possible that the school-level R^2 could be higher than assumed, but we erred on the side of being more conservative in the power calculation to ensure sufficient power to detect hypothesized effects.

[55] Two classrooms per school were sampled for three reasons. First, we wanted school-level estimates to be based on a sample size of more than one classroom. Second, we wanted to allocate resources to maximize the number of schools more than the number of classrooms per school. Third, although some schools in the Mississippi Delta have many kindergarten classrooms, some schools have only two classrooms, and we did not want to eliminate those schools from the sample.

[56] Based on the assumption of a 20% student attrition rate, eight students per classroom were assumed in the statistical power analysis.

ACTUAL MINIMUM DETECTABLE EFFECT SIZES FOR STUDENTS' EXPRESSIVE VOCABULARY

The actual analytic sample size at the end of kindergarten was 64 schools, 128 classrooms, and 1,296 students (or 10 students per classroom). Once the data were collected, the actual minimum effect size that could be detected with 80% power was calculated using the standard error of the estimated treatment effect from the fitted impact models. The following equation was used to calculate the actual minimum detectable effect size in the sample for the Expressive Vocabulary Test–2nd Edition (EVT–2) posttest:

(B2)

$$MDES(\hat{\beta}_1\ treatment) = \frac{Factor(\alpha, \beta, df) * S.E.(\hat{\beta}_1\ treatment)}{SD_{CONTROL}}$$

where $MDES(\hat{\beta}_1\ treatment)$ is the estimated minimum detectable effect size for the treatment impact; $Factor(\alpha, \beta, df)$ is a constant that is a function of the significance level (α), statistical power (β), and the number of degrees of freedom (*df*); and $SD_{CONTROL}$ is the standard deviation of the EVT–2 posttest in the control group, in the units of the EVT–2 posttest score. This calculation indicated that a standardized effect size of 0.144 standard deviation could be detected for the EVT–2.

Because the minimum detectable effect size was lower than anticipated, the assumptions were compared with the observed data for the EVT–2 (Table B2). Appendix K presents a multilevel model estimated to test the impact of K-PAVE on students' expressive vocabulary, including a list of the student and school-level covariates.

Table B2. Comparison of assumed and observed factors related to minimum detectable effect size for the Expressive Vocabulary Test–2 posttest

	Assumed	Observed
Proportion of variation between schools	.10–.15	.11
Proportion of variation between classrooms	.05	.01
Student attrition rate	.20	.07
School-level R^2	.50	.99

The observed between-school variation in EVT–2 posttest scores was within the assumed range; however, for the other factors, the observed data indicated that the assumptions about statistical power were too conservative. The proportion of variation between classrooms was lower than assumed (.01 compared with .05). Most notable, however, was the difference between the observed school-level R^2 of .99 and the assumed R^2 of .50. The school and student covariates in the impact model accounted for nearly all the between-school variation in students' EVT–2 posttest standard scores.[57]

[57] The student-level covariates in the model were baseline EVT–2 score, gender, race/ethnicity, eligibility for free or reduced-price meals, and special education status (having an individualized education plan). The school-level covariates in the model were previous reading initiative (Reading First, a state initiative, or other), state rating of school achievement level index (created based on student performance on the Mississippi state accountability test administered to students in grades 3 and higher), percentage of students in the school who are African American,

The higher school-level R^2 and the lower classroom-level variation contributed to the study's power to detect a lower effect size than had been assumed. Although the study was thought to have 80% power to detect an effect size of 0.26 to 0.28 or higher, the actual minimum detectable effect size for EVT–2 was 0.149, and the observed standardized effect size of 0.141 was found to be statistically significant (see Table 4.1 in Chapter 4).

STATISTICAL POWER FOR DETECTING IMPACTS ON CLASSROOM INSTRUCTION

Given the recruitment target of 60–70 schools, a statistical power analysis was conducted to determine the minimum detectable effect sizes for classroom instruction outcomes (table B3).[58] The following equation (Schochet 2008a) was estimated with terms defined as described for equation B2:

(B3)

$$MDES(\hat{\beta}_1\ treatment) = Factor(\alpha, \beta, df) * \frac{\sqrt{\frac{\sigma^2_{school}(1-R^2_{school})}{sp(1-p)} + \frac{\sigma^2_{class}(1-R^2_{class})}{(sp(1-p))c}}}{\sigma}$$

with the following assumptions:

- A two-tailed test of significance will be conducted at the .05 level.
- The proportion of total variation in classroom outcomes at the school level is .15.[59]
- The school-level R^2 and classroom-level R^2 were both 0, in order to be conservative in the power calculation.
- The desired power to detect effects is .80.

With a sample of 60–70 schools, the minimum detectable effect size was estimated to be 0.51–0.56.60 Although this range is substantially larger than those for impacts on students, this range is appropriate for examining impacts on classrooms. Because the intervention is intended to directly impact instructional practices, larger effects are expected on classrooms than on students.

---

percentage of students in the school who are eligible for free or reduced-price meals, locale (rural, small town, large town/fringe of city), and location in or out of the Delta region.

[58] The statistical power analysis for classroom instruction outcomes was not conducted to determine the sample size but rather to estimate minimum detectable effect sizes, given the targeted sample size. The study was designed to detect impacts on students.

[59] We intended to make a conservative assumption about the intraclass correlation, so we assumed one at the upper range of the assumption in the statistical power analysis for detecting impacts on students. We were unaware of empirical evidence for guiding this assumption.

[60] Because there is higher measurement error in the classroom outcomes than is typically the case for standardized student outcome measures, our ability to detect effects on classroom instruction is more limited than these calculations suggest. Even so, the number of schools in the study need not be increased, since the project's primary purpose is to measure impacts on students in a sufficiently precise manner.

Table B3. Power analysis summary: minimum detectable effect sizes for classroom instruction outcomes, by number of schools

Intraclass correlation	70 schools	65 schools	60 schools
.15	0.51	0.53	0.56

The results of the a priori statistical power analysis and the known analytic sample of 128 classrooms and 64 schools suggested that the minimum detectable effect size for classroom instruction outcomes with 80% power was 0.53.

ACTUAL MINIMUM DETECTABLE EFFECT SIZES FOR CLASSROOM INSTRUCTION OUTCOMES

Once the data were collected, the actual minimum effect size that could be detected with 80% power was calculated using the standard error of the estimated treatment effect from the impact models for each of the classroom instruction outcomes. Just as with the EVT–2, the actual minimum detectable effect size was calculated using equation B2. Table B4 shows the minimum detectable effect sizes in the observed data for each of the four classroom instruction outcomes.

Table B4. Minimum detectable effect sizes and estimated impacts in the observed sample for classroom instruction outcomes

	Minimum detectable effect size	Estimated treatment impact
Vocabulary and comprehension support	0.67	0.82***
Instructional support	0.77	0.47
Emotional support	0.74	0.20
Time on non–vocabulary literacy instruction	0.61	–0.03

*** $p < .001$.

For impacts on classroom instruction, the minimum detectable effect sizes were not as low as anticipated. Table B5 compares the assumptions with the observed data for classroom instruction.

Table B5. Comparison of assumed and observed factors related to minimum detectable effect sizes for classroom instruction outcomes

	Intraclass correlation		School-level R^2	
	Assumed	Observed	Assumed	Observed
Vocabulary and comprehension support	.15	.41	.00	.10
Instructional support	.15	.66	.00	.00
Emotional support	.15	.33	.00	.00
Time on non–vocabulary literacy instruction	.15	.60	.00	.06

Although the observed school-level R^2 values were close to the assumption, the observed intraclass correlations were higher than assumed, which contributed to higher minimum detectable effect sizes for classroom instruction than anticipated.[61] That is, more of the variation

[61] The finding that the school-level variation in classroom instruction measures is higher than typically found for student outcomes may not be unusual. Schochet (2009) reviewed studies testing impacts on classroom instruction

in classroom instruction than assumed was between schools rather than within schools, resulting in less power than estimated. Although the study was thought to have 80% power to detect effects on classroom instruction of 0.53 standard deviation or higher, the minimum detectable effect sizes ranged from 0.63 to 0.78.

---

outcomes and found that intraclass correlations ranging from 0.20 to 0.33 for a study evaluating reading comprehension curricula in elementary school.

APPENDIX C. RANDOM ASSIGNMENT

MATCHING OF SCHOOLS WITHIN BLOCKS FOR RANDOM ASSIGNMENT

Seventy schools were blocked into three groups based on previous participation in reading initiatives: Reading First (n = 17 schools), Mississippi state reading initiative (n = 7 schools), or local district initiative or no reading initiative (n = 46 schools). Within these three blocks, schools were matched based on a set of school characteristics (see Table 2.1 in Chapter 2 for sample distributions):

- School performance classification[62]
    - Low or underperforming
    - Successful
    - Exemplary or superior
- Percentage of students receiving free or reduced-price meals
    - 96%–100%
    - 90%–95%
    - 70%–89%
    - Less than 70%
- Percentage of students that are African American
    - 96%–100%
    - 81%–95%
    - 21%–80%
    - Less than 20%
- Locale
    - Rural
    - Small town
    - Large town or fringe of city
- Region
    - Delta
    - Outside

School characteristics were ordered such that those hypothesized to be more strongly associated with student outcomes were prioritized in the matching process. [63] Within reading initiative

[62] The School Performance Classification is an annual classification based on students' performance on the state accountability test (Mississippi Curriculum Test) administered to students in grades 3 and higher. Classifications include: low performing, underperforming, successful, exemplary, and superior.

blocks, schools were sorted by all five characteristics in the order listed above—from school performance classification to region. Specifically, within the three reading initiative blocks, schools were sorted into the three school performance classifications from lowest performing to highest performing (that is, low or underperforming, successful, and exemplary or superior). Within each school performance classification, schools were sorted into four categories from highest to lowest percentage of students receiving free or reduced-price meals. Within each of these four categories, schools were sorted into four categories from the highest to lowest percentage of African American students. Within each category regarding the percentage of African American students in the school, schools were sorted into three categories from most to least rural. Within each locale category, schools were sorted into two regions–Delta and outside.

Once schools were ordered, the first school within each reading initiative block was randomly assigned to either the intervention or control condition. The next school on the ordered list—the first school's match—was assigned to the other condition. The third school on the list was assigned to the first condition, and the fourth school—the third school's match—was assigned to the same condition as the second school. The assignment of schools alternated between each condition until all schools were assigned to either the intervention or control condition; 35 schools assigned to the K-PAVE intervention and 35 schools assigned to the control condition.

By sorting and then assigning schools in this manner, each school was matched to the school most similar to it in terms of these five school characteristics. However, characteristics that were given lower priority in the matching process (used later in the sorting order) had less influence on the matches than characteristics that were given higher priority. A hypothetical example can be used to illustrate the relative influence of school characteristics in the matching process. For example, if within the Reading First block, there are six schools classified as low or underperforming, they are grouped together. Next, these six schools would be ordered based on the percentage of students eligible for free or reduced-price meals, four of which have 96% or more students eligible. In two of those schools, 96% or more students are African American. By the time the last two characteristics—locale and region—are factored in to the matching process in this example, there are only two Reading First schools that have been grouped together based on the three higher priority school characteristics—school performance classification, eligibility for free or reduced-price meals, and percentage of African American students. Regardless of the values of locale and region, these two schools cannot be sorted any further; they have already been matched with each other. One school will be randomly assigned to the intervention condition, regardless of the values for locale and region.

PROCESS OF RANDOM ASSIGNMENT: SEQUENCE GENERATION

Random assignment was conducted by the evaluation team. The evaluation team was independent of the intervention team, which was responsible for intervention training and support, and was independent of the school districts and schools, which were responsible for intervention delivery.

The process of generating a random sequence of numbers in order to randomly assign schools to either the K-PAVE intervention condition or the control condition was conducted using SAS

[63] We needed to rely on hypotheses because at the time of random assignment we did not have data with which to estimate the relationship between each school characteristic and the student outcomes examined in the study.

computer software (version 9.2 for Windows). As noted above, schools were ordered within reading initiative blocks based on a set of school characteristics. With the ordered dataset for any block, we used the RANUNI function in SAS, which returns a number that is generated from a uniform distribution on the interval (0, 1) using a prime modulus multiplication generator with modulus 2^{31} and multiplier 397204094 (Fishman & Moore, 1982). Specifically, we used the following SAS code:

Treatment = INT ((RANUNI (0) * 2) + 1).

The RANUNI function requires a seed value, which is a numeric constant that provides an initial starting point for generating a stream of random numbers. By specifying zero as the seed, as we did, the computer clock initializes the stream. Including a multiplier (2, in this case) changes the length of the interval, and adding a constant (1, in this case) moves the interval. In this case the RANUNI function returns a number that is generated from a uniform distribution on the interval (1, 2). The INT function truncates the decimal portion of the number generated by the RANUNI function, thus yielding an integer, either 1 or 2, with a 50 percent probability. For a value of 2, the school was assigned to the K-PAVE intervention; for a value of 1, the school was assigned to the control condition.

However, because schools were ordered based on school characteristics within reading initiative blocks, we used the random number generated by the RANUNI function *only for the first school within the block*. Once the assignment of the first school was complete, the next school—its matched-pair mate—was assigned to the other condition. The remaining schools on the ordered list were assigned to their condition in an alternating sequence, as described above. In other words, the assignment of schools on the ordered list alternated between each condition until all schools within a block were assigned to either the intervention or control condition.

Table C.1 illustrates the alternating sequence for a hypothetical list of schools within the Reading First block. Based on the RANUNI function in SAS, the first school in the block was randomly assigned to the K-PAVE intervention. The remaining schools were alternately assigned to the control and intervention conditions.

Table C1 Random assignment for a hypothetical list of Reading First Schools, ordered based on school characteristics

School ID	Reading initiative	School performance classification	Percent eligible for free/ reduced price lunch	Percent African American	Locale	Region	Treatment
1	Reading First	Under-performing	96%-100%	96%-100	Rural	Delta	1
2	Reading First	Under-performing	96%-100%	96%-100	Small town	Outside	0
3	Reading First	Under-performing	90%-95%	81%-95%	Rural	Delta	1
4	Reading First	Under-performing	70%-89%	81%-95%	Small town	Delta	0
5	Reading First	Under-performing	70%-89%	21%-80%	Large town	Outside	1
6	Reading First	Successful	96%-100%	81%-95%	Rural	Delta	0
7	Reading First	Successful	90%-95%	96%-100	Rural	Outside	1
8	Reading First	Successful	90%-95%	81%-95%	Small town	Delta	0
9	Reading First	Successful	Less than 70%	96%-100	Large town	Delta	1
10	Reading First	Exemplary/Superior	90%-95%	81%-95%	Rural	Delta	0
11	Reading First	Exemplary/Superior	90%-95%	21%-80%	Large town	Delta	1
12	Reading First	Exemplary/Superior	Less than 70%	21%-80%	Small town	Delta	0
13	Reading First	na	96%-100%	96%-100	Rural	Outside	1

na is not applicable.

Allocation concealment

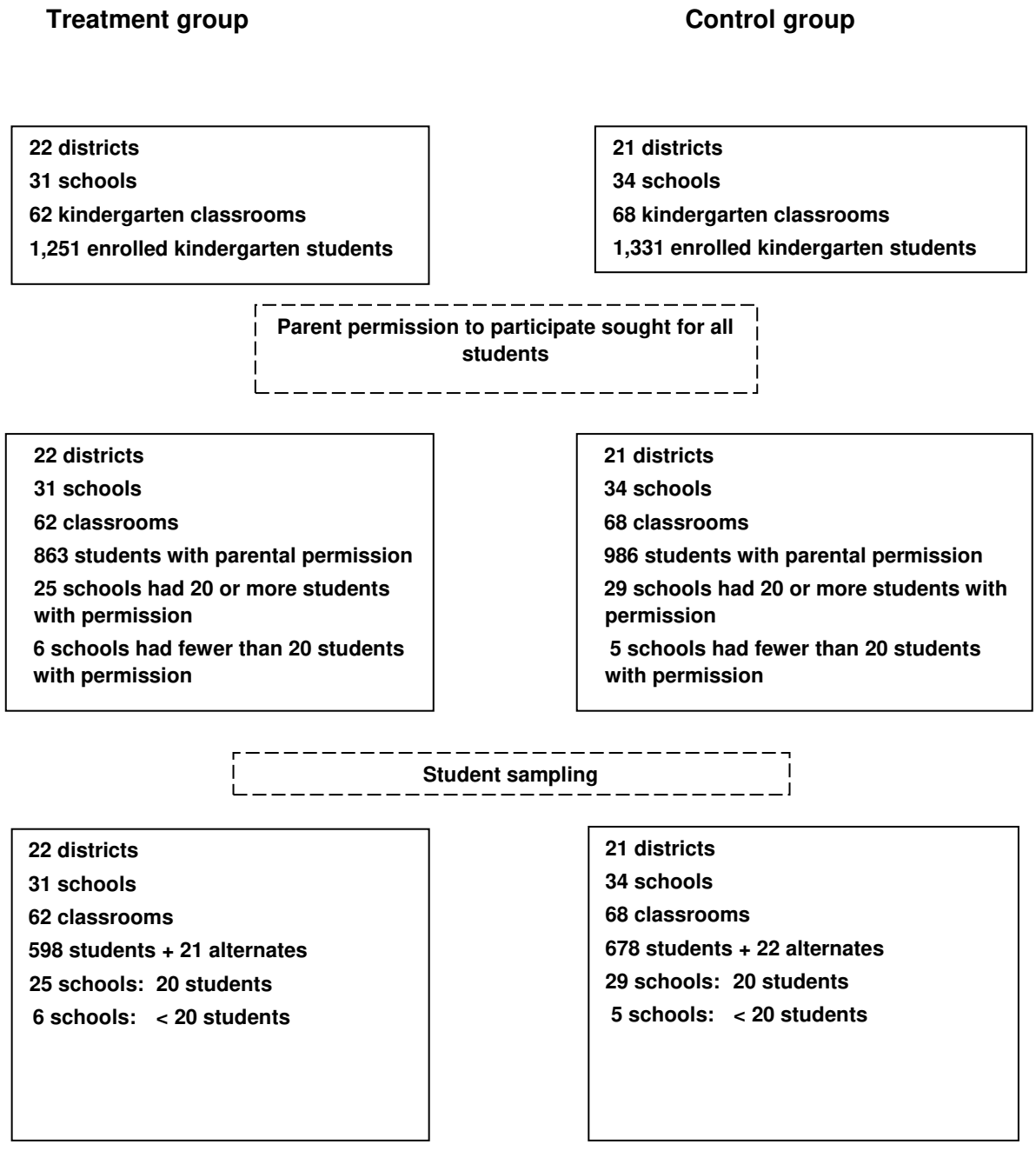
For random assignment to be successful, the process by which study units are randomly allocated to the intervention or control condition must be concealed from *both* the study participants and the evaluators as the allocation takes place (Forder, Gebski, & Keech, 2005). If either study participants or the evaluators are able to influence the random allocation process in progress, then the equivalence of the intervention and control groups may be compromised. If the evaluator has influence over which units are assigned to a given condition, then the process of assignment is not random. Similarly, if participants' willingness to enroll in the study is influenced by whether they are assigned to the intervention or control condition, then self-selection bias creating treatment/control differences is introduced.

In this study, allocation remained concealed from both evaluators and study participants during the randomization process. As described above, the process for blocking schools based on reading initiative and ordering schools within blocks based on school characteristics was defined prior to initiating the random assignment process and was not influenced by any examination of school characteristics data. In addition, once schools were ordered through an automated computer sorting of the schools, an automated computer process for generating the random allocation was used in order to ensure that there was a 50 percent chance of assignment to either the intervention or control condition. The evaluator had no control in determining the allocation of schools apart from initiation of randomization by the computer.

For study participants, the decision to enroll in the study was not influenced by random assignment (i.e., by whether a school would receive the intervention in 2008-09 or a year later). Districts, schools, and teachers were not notified of random assignment status until all required paperwork for study participation was submitted. Concealing random assignment status from schools until enrollment was completed guards against the introduction of selection bias as part of the study participation process.

APPENDIX D. RECRUITMENT AND RANDOM SELECTION OF THE STUDENT SAMPLE

Figure D1. Recruitment and random selection of the student sample



APPENDIX E. COMPARISON OF STUDENTS MISSING AND NOT MISSING BASELINE ASSESSMENT

Table E1. Characteristics of students missing and not missing baseline assessment

	Missing baseline assessment *n*=46	Not missing *n*=1250	Test of difference[a]
Student gender			
Female	47.4%	50.0%	t = -0.32, p=.75
Male	52.6%	50.0%	
Student race/ethnicity			
African American	74.4%	85.3%	t = 2.49, p=.007
Other	25.6%	14.7%	
Eligibility for free or reduced-price meals			
Eligible /free or reduced-price	92.1%	92.9%	t=0.37, p=.71
Not eligible	7.9%	7.1%	
Has an Individualized Education Plan			
Yes	17.5%	7.9%	t= 1.90, p=.06
No	82.5%	92.1%	
Age at posttest			
Mean	6 yrs, 2.8 mo	6 yrs, 1.9 mo	t= 1.08, p=.28
Standard deviation	4.8 mo	4.7 mo	

Note: The distribution of student characteristics for cases with missing data is assumed to be the same as for cases with non-missing data. The rate of missing data ranges from 0.1% to 21.7%.

a. Differences in student characteristics between the missing and nonmissing groups were tested using a model with a three-level error structure to account for the nesting of students within classrooms and classrooms within schools; no covariates were included in the model other than presence or absence of a baseline assessment. Although four of the student characteristics are dichotomous (gender, race/ethnicity, eligibility for free or reduced-price meals, and special education status), tests were conducted using a linear model, which yielded a *t*-test of the mean difference between students with and without a baseline assessment (where the mean for each characteristic equals the percentage of students in each group that are female, African American, eligible for free or reduced-price meals, or in special education, respectively). A chi-square test, which is usually used to test group differences in categorical variables, would not take account of the multilevel structure of the data.

APPENDIX F. CLASSROOM OBSERVATION MEASURES FOR IMPACT EVALUATION

This appendix discusses three classroom observation measures for impact evaluation: Classroom Assessment Scoring System (CLASS), Read Aloud Profile–Kindergarten (RAP–K), and Vocabulary Record. In addition, the last section describes the vocabulary and comprehension support composite, which is created from two RAP-K variables and two Vocabulary Record variables.

CLASSROOM ASSESSMENT SCORING SYSTEM

The CLASS is a time-sampling observation tool used to rate the quality of interactions and instruction in kindergarten to grade 3 classrooms. The CLASS has been widely used in evaluations of education interventions, has established interrater and internal consistency reliability, and has established construct and predictive validity (Pianta, La Paro, & Hamre 2008). The CLASS shows statistically significant correlations with the Early Childhood Environment Rating Scale–Revised interactions factor (r = 0.45–0.63 for various CLASS domains). In addition, in the National Center for Early Development and Learning Multi-State Pre-Kindergarten Study, children in preschool classrooms with higher CLASS scores demonstrated higher performance on vocabulary, prereading, and mathematics outcomes at the end of preschool and showed greater gains during preschool on these outcomes as well as fewer behavior problems, than did children in classrooms with lower CLASS scores (Howes et al. 2008; Mashburn et al. 2008).

The CLASS is a theoretically and empirically based classroom observation tool. As stated in the CLASS manual (Pianta, LaParo, & Hamre 2008, p. 1):

> The CLASS dimensions are based on developmental theory and research suggesting that interactions between students and adults are the primary mechanism of student development and learning (Greenberg, Domitrovich, & Bumbarger 2001; Hamre & Pianta 2007; Morrison & Connor 2002; Pianta 2006; Rutter & Maughan 2002). The CLASS dimensions are based solely on interactions between teachers and students in classrooms; this system does not evaluate the presence of materials, the physical environment or safety, or the adoption of a specific curriculum. … The CLASS focuses on interactions between teachers and students and what teachers *do* with the materials they have [emphasis in original].

The CLASS measures three domains of classroom quality: instructional support, emotional support, and classroom organization.

Instructional support

The instructional support domain examines the extent to which interactions effectively support cognitive and language development and provide opportunities for children to gain useable knowledge—learning how facts are interconnected and organized—and develop metacogntive skills. The dimensions of instructional support are:

- *Concept development.* How teachers use instructional discussions and activities to promote students' higher order thinking skills in contrast to a focus on rote instruction:
    - Analysis and reasoning
    - Creating
    - Integration
    - Connections to the real world
- Quality of feedback. How teachers extend students' learning through their responses to students' ideas, comments, and work:
    - Scaffolding
    - Feedback loops
    - Prompting through processes
    - Providing information
    - Encouragement and affirmation
- *Language modeling.* The extent to which teachers facilitate and encourage students' language:
    - Frequent conversation
    - Open-ended questions
    - Repetition and extension
    - Self- and parallel-talk
    - Advanced language

Emotional support

Teachers' ability to support children's social and emotional functioning is considered central to effective classroom practice, as children who are motivated and connected to others in the early years of schooling are much more likely to have positive social and academic outcomes. The dimensions of emotional support are:

- *Positive climate.* The emotional connection, respect, enjoyment demonstrated between teachers and students and among students:
    - Relationships
    - Positive affect
    - Positive communication
    - Respect
- Negative climate. The level of expressed negativity such as anger, hostility, or aggression exhibited by teachers and or students in the classroom:
    - Negative affect

    - Punitive control
    - Sarcasm/disrespect
    - Severe negativity
- *Teacher sensitivity.* Teachers' awareness of and responsivity to students' academic and emotional concerns:
    - Awareness
    - Responsiveness
    - Addresses problems
    - Student comfort
- *Regard for student perspectives.* The degree to which teachers' interactions with students and classroom activities place an emphasis on students' interests, motivations, and points of view:
    - Flexibility and student focus
    - Support for autonomy and leadership
    - Student expression
    - Restriction of movement

Classroom organization

The classroom organization domain assumes that students are able to monitor, regulate, and control their learning, motivation, and behavior in well regulated classroom environments. The dimensions of classroom-level regulation measured by the CLASS are:

- *Behavior management.* How effectively teachers monitor, prevent, and redirect behavior:
    - Clear behavior expectations
    - Proactive
    - Redirection of misbehavior
    - Student behavior
- *Productivity.* How well the classroom runs with respect to routines and the degree to which teachers organize activities and directions so that maximum time can be spent in learning activities:
    - Maximizing learning time
    - Routines
    - Transitions
    - Preparation
- Instructional learning formats. How teachers facilitate activities and provide interesting materials so that students are engaged and learning opportunities are maximized:

- Effective facilitation
- Variety of modalities and materials
- Student interest
- Clarity of learning objectives

The CLASS is completed based on at least two hours of observation of a classroom. Coding involves 30-minute cycles (20 minutes for observing and 10 minutes for rating each dimension). The 10 dimensions described above are rated on a seven-point Likert scale. Ratings of the 10 dimensions are used to compute the three domain scores and a total score, in accordance with guidelines outlined by the instrument developers.

READ ALOUD PROFILE–KINDERGARTEN

The RAP–K was adapted from the Read Aloud Profile–Revised (RAP) instrument from the Observation Measures of Language and Literacy Instruction (Goodson, Layzer, Smith, & Rimdzius 2004). The original instrument was adapted to focus during book readings primarily on teachers' comprehension support statements and questions, teachers' open-ended questions, and teachers' emphasis on word meanings. For this study the instrument was modified to eliminate the focus on other aspects of literacy, such as book concepts and print concepts (including letter names, letter sounds, decoding, punctuation, and spelling). The process of adapting the RAP into the RAP–K involved multiple iterations during which coders jointly and then later independently coded a series of video recordings of teacher-child book readings. Once a near-final version of the instrument was completed, the instrument was pilot-tested in five kindergarten classrooms in May 2008. Slight modifications were made to the format of the coding form based on the pilot, but the coding rules remained unchanged.

The reading instructional strategies captured by the RAP-K include how the reader reads the book and how the reader interacts orally with children during the text reading to build their comprehension and vocabulary. This measure focuses on reading aloud because of the widespread recognition that reading aloud to children is one of the "most important activities for building the knowledge required for eventual success in reading" (Anderson, Hiebert, Scott, & Wilkinson, 1985, p. 23). The RAP-K is designed to measure interactive instructional practices in reading aloud (sometimes called "dialogic reading") that research has shown to promote children's comprehension and higher order thinking abilities (see review of shared reading interventions in Chapter 4 of NELP 2009).[64]

The RAP-K focuses on the behavior of the reader during the read-aloud and provides information on the characteristics of the book being read. The RAP-K describes two aspects of the reader's strategies during the read-aloud: the reader's use of comprehension supports before, during, and after the text reading and the reader's use of higher order, cognitively challenging questions (Box F1).

During the classroom observation, the RAP-K was coded the first time that a teacher read aloud to a group of students, at which time the observer stopped coding the CLASS until the read aloud

[64] The research on effective practices for reading with children is based primarily on a teacher or parent reading with an individual child (see chapter 4 of NELP 2009). The RAP-K is based on the assumption that many of the same practices that have been shown to be effective in one-on-one contexts may also be effective with groups of children.

ended. Throughout the entire book reading, from the time the teacher announces that she is going to read a book until any post-reading discussion of the book is concluded, the observer documents the number of comprehension supports provided during reading, including providing background information related to the book, making connections to children's experiences, and asking concrete or factual questions to clarify meaning, and the number of higher order questions asked during reading, including asking students to analyze, explain, predict, imagine, make inferences, or generate hypotheses. If the teacher does not read aloud to a group of students during the classroom observation, then all behaviors are coded as occurring zero times; they are not coded as missing. If the teacher reads aloud to a group of students more than one time during the observation, the observer continues coding the CLASS for all additional instances of reading aloud; the RAP-K is coded only the first time that the teacher reads aloud.

Box F1. Read Aloud Profile–Kindergarten coding form

CLASS cycle interrupted: _____ Start Time __ : am/pm End time __ : ___ am/pm	Book Title:	# students	Reader: Teacher Assistant (circle one)
Book Type: ❑ Storybook ❑ Expository	Words/page: ❑ 0 ❑ 1 ❑ 2-10 ❑ 10+	Special Events: ❑ No RAP-K codes	❑ Book not started ❑ Book not finished

	1 Comprehension support			Total
a	Provides additional information related to text; clarifies meaning; expands on text (non-vocab)	C		
b	Relates text to class activities; reminds children of same/similar book read before	C or Q	❑	
c	Narrates/tells the story in advance of reading	C	❑	
d	Talks about events and/or features to listen, look for in the story/book	C	❑	
e	Connects book to children's personal experiences with comments or questions	C or Q		
f	Comments on picture/asks question with known answer about picture	C or Q		
h	Asks question with known answer	Q		
i	Summarizes story (or asks child); acts out story (or asks child)	C or Q	❑	
x	Questions or comments about print, spelling, letters, sounds, punctuation	C or Q	❑	
2	**Higher-order questions**			
a	Prediction (what's going to happen in story); Analysis/Inference/Explanation (Why? How?)	Q		
b	Imagine things, events or situations outside children's experience (possible or fantastical)	Q		

3. Vocabulary support	Asks Ss for meaning	Defin/syn/ex	Picture/Demo	Contrast
a.	❑	❑	❑	❑
b.	❑	❑	❑	❑
c.	❑	❑	❑	❑
d.	❑	❑	❑	❑
e.	❑	❑	❑	❑
f.	❑	❑	❑	❑
g.	❑	❑	❑	❑
h.	❑	❑	❑	❑
i.	❑	❑	❑	❑
j.	❑	❑	❑	❑
k.	❑	❑	❑	❑
l.	❑	❑	❑	❑
m.	❑	❑	❑	❑
n.	❑	❑	❑	❑

VOCABULARY RECORD

The Vocabulary Record was adapted from the vocabulary component of the Instructional Practice in Reading Inventory (Smith, Dwyer, Dixon, Schimmenti, Boulay, Khalil, Blocklin, & Gamse 2005). Although the Instructional Practice in Reading Inventory was designed to document a broad array of literacy instruction in kindergarten to grade 3 classrooms, the Vocabulary Record focuses exclusively on teachers' attention to word meaning. Strategies for communicating word meaning were discussed and combined into like categories. For example, providing word meaning through a definition, an example, or a synonym were all considered a single strategy, while using a picture or physical demonstration of word meaning was considered a distinct strategy. Stating what a word is not or does not mean was considered a third distinct

strategy for communicating word meaning. The process of adapting the Instructional Practice in Reading Inventory into the Vocabulary Record involved multiple iterations, involving both piloting early versions in kindergarten classrooms and coding of video recordings of teachers. The initial piloting and video coding informed the instrument development. A second pilot test in five kindergarten classrooms was conducted during which two raters independently coded the Vocabulary Record. One modification was made to the instrument based on the pilot: a code for the teacher asking the student to define a word was added.

The Vocabulary Record is completed during the CLASS observation cycle and during the read-aloud coded with the RAP-K. The observer records all words defined by the teacher or assistant teacher and all words that the teacher or assistant teacher ask a student to define (Box F2). For each word entered on the Vocabulary Record, the observer indicated how meaning was elaborated by marking all the check boxes that apply for that word:

- Asks student for meaning
- Definition, example, or synonym
- Picture or demonstration
- Contrast

For the current study, this instrument yielded two variables. The total number of words documented during the read aloud was tallied for a measure of the number of words introduced during the book reading. The total number of words documented during each CLASS cycle was tallied and averaged (i.e., divided by the number of CLASS cycles) for a measure of the average number of words introduced during other instructional time.

CREATION OF VOCABULARY AND COMPREHENSION SUPPORT COMPOSITE

A composite measure of vocabulary and comprehension support was created from the two variables from the Vocabulary Record (number of words introduced during the read-aloud and the average number of words introduced during other instructional time) and the two variables created from the RAP-K (number of comprehension supports during reading and number of higher order questions during reading). The four variables were standardized to ensure that each would be equally weighted in the composite total. The standardized variables were summed, and the total composite score was then standardized to a mean of 0 and a standard deviation of 1. The standardized composite score units are standard deviation units, such that a score of 0 indicates an average amount of vocabulary and comprehension support was provided in the classrooms and a score of 1 indicates that the amount of vocabulary and comprehension support is one standard deviation higher than average. A positive score indicates that a higher than average amount of vocabulary and comprehension support is provided, while a negative score indicates that a lower than average amount of comprehension support is provided.

Cronbach's alpha for the vocabulary and comprehension support composite is 0.62. Each variable created from the book reading—comprehension support, higher order questions, and number of vocabulary words—has a correlation, with the composite ranging from .45 to .55. Removing any of the variables from the composite reduces the internal consistency of the composite (that is, Cronbach's alpha becomes smaller). The correlation between the average number of words introduced during other instructional time—collected throughout the rest of the

three-hour observation other than the book reading coded with the RAP–K—and the composite variable was .16. Removing this variable from the composite would increase the internal consistency (that is, Cronbach's alpha would be 0.71). However, all four variables were retained in the composite for analysis. The composite was determined based on the theorized relation among the four variables as part of a single domain prior to any examination of the relationships observed in the sample data. An empirically defined composite that may have been overly influenced by sampling variation was to be avoided.

Box F2. Vocabulary Record coding form

CLASS cycle # 1											
3	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**	**3**	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

CLASS cycle # 2											
3	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**	**3**	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

CLASS cycle # 3											
3	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast	3	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

CLASS cycle # 4											
3	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast	3	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

CLASS cycle # 5 **3 Vocabulary support**		**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**	**3**	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

CLASS cycle # 6 **3 Vocabulary support**		**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**	**3**	**Vocabulary support**	**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**
a		❑	❑	❑	❑	h		❑	❑	❑	❑
b		❑	❑	❑	❑	i		❑	❑	❑	❑
c		❑	❑	❑	❑	j		❑	❑	❑	❑
d		❑	❑	❑	❑	k		❑	❑	❑	❑
e		❑	❑	❑	❑	l		❑	❑	❑	❑
f		❑	❑	❑	❑	m		❑	❑	❑	❑
g		❑	❑	❑	❑	n		❑	❑	❑	❑

APPENDIX G. TEACHER SURVEY

OMB Number: 1850-0846
Expiration Date: 12/31/2010

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

Teacher Name_____School Name_____

Teacher Number _____ School Number_____

Birth Date (Month, Day, Year): ___/_____/____

1. What is your gender? O Female O Male

2. What is your race? (Select one or more)

O African American
O White
O Asian
O
O American Indian
O Pacific Islander/Hawaiian
O Multiracial
O Unknown

3. What is your ethnicity? O Hispanic O Non-Hispanic O Unknown

4. EDUCATIONAL BACKGROUND AND PROFESSIONAL EXPERIENCE
Please check and complete for all that apply.

Education	Major	Year Completed
O High School	_____	_____
O GED	_____	_____
O Non-degree program (e.g. Montessori, CDA)	_____	_____
O Some college/university	_____	_____
O Bachelor's degree	_____	_____
O Some graduate level classes	_____	_____
O Master's degree	_____	_____
O Education Specialist	_____	_____
O Doctorate	_____	_____

5. Please check all areas in which you have a current teaching certificate.

O Early Childhood
O Middle Childhood
O Secondary
O ESOL
O Special Education
O Gifted/Talented
O Administration
O Reading
O Other _____

6. Do you have any other special training? O Yes O No

Please describe. _____

We would like to learn about teachers' experiences collaborating with other teachers in their schools. Please think about both formal activities at your school intended to encourage collaboration and informal conversations you have with other teachers.

7. Not including the current school year and not including student teaching, how many years have you been a teacher? *If this is your first year teaching, answer "zero."* _____ years

8. Not including the current school year and not including student teaching, how many years have you been teaching kindergarten?

   If this is your first year teaching, answer "zero." _____ years

9. Not including the current school year and not including student teaching, how many years have you taught in your current school? *If this is your first year in this school, answer "zero."* _____ years

10. Some teachers work independently while other teachers prefer to get input from other teachers. Would you say you get…

    O No input
    O Minimal input
    O Moderate input
    O A great deal of input

11. How comfortable are you receiving advice from other teachers?

    O Not at all comfortable
    O Slightly comfortable
    O Moderately comfortable
    O Completely comfortable

12. How comfortable are you offering advice to other teachers?

    O Not at all comfortable
    O Slightly comfortable
    O Moderately comfortable
    O Completely comfortable

13. How supportive are other teachers at your school when you need help or advice with teaching?

    O Virtually no teachers are supportive
    O Some teachers are supportive, but a majority are not
    O A majority of teachers are supportive, but some are not
    O Nearly every teacher is supportive

14. How receptive are other teachers at your school when you offer help or advice with teaching?

    O Virtually no teachers are supportive
    O Some teachers are receptive, but a majority are not
    O A majority of teachers are receptive, but some are not

O Nearly every teacher is receptive

15. In general, how often do you participate in any organized group activities or meetings involving other teachers at your school…

…that primarily focus on administrative issues, such as schedules, upcoming events, and teachers work assignments?

Number of times: _____ O per week
O per month
O per year

…that primarily focus on issues pertaining to student instruction/behavior?

Number of times: _____ O per week
O per month
O per year

16. Think of changes that you have made over the past year that were due to a suggestion from another teacher in your school OR due to your having observed another teacher in your school.

Do NOT include changes that were due to a principal, or to someone outside of your school, that you were required to make, or that occurred as a regular part of the school calendar (for example, changes that always occur when switching from fall to spring semesters).

Changes in… *Mark all that apply*
…classroom materials that you use

O Handouts
O Books
O Hands-on learning materials
O Computer software
O Assessments (tests)
O Behavior charts
O Parent communication product (for example, daily reports)
O Other *(please describe)*_____
_____

O how you teach lessons that you've taught in the past
O curriculum that involve teaching new lessons
O the homework you assign to students
O how you handle behavior problems involving an individual student
O your overall approach to managing student behavior in your class
O classroom management unrelated to discipline
O strategies for communicating with parents
O the classroom setting (physical environment)
O your own understanding of materials/procedures that you currently use
O your own understanding of the *content* of what you teach
O your approach to teaching specific groups of students (for example, students who are less proficient in English than they are in another language)

- your approach to any aspect of extra-curricular activities that you might be involved with (for example, coaching, tutoring or helping in an after school program)

APPENDIX H. K-PAVE FIDELITY OBSERVER HANDBOOK AND TRAINING FIDELITY CHECKLIST

KINDERGARTEN PAVED FOR SUCCESS (K-PAVE) FIDELITY OBSERVATIONS

Scheduling observations/protocol

Observations of the intervention classrooms will be conducted within one week of the training. As Bowman, Donovan, and Burns (2001) suggest "effective in-service education must be intensive and continuous, with opportunities to apply knowledge and receive individualized feedback and mentoring in order to support improved teaching practices and positive outcomes for children" (p. 276). The intervention designers, their assistants, and REL-SE staff will visit each K-PAVE focal teacher's classroom to assess the degree to which teachers are implementing the intervention as taught in the training. A 90-minute observation will be conducted in each focal teacher's class. So that the observation is conducted during the typical literacy block, focal teachers will provide the designers/fidelity observers with a class schedule that details when the literacy block of instruction is typically taught. The observations will be scheduled ahead of time to allow the fidelity observers to evaluate the degree to which teachers have learned the vocabulary intervention components. Teachers may be asked to carry out their literacy block at a time other than their typical time to allow observers to maximize the number of observations that can be carried out at a given school.

Forms used

There are two forms that need to be completed after each 90-minute observation–the *Training Fidelity Checklist* and the *Teacher Observation Follow-Up Meeting Protocol*. This handbook provides details on how to complete each. Observers will need to use their field notes to complete these forms. **These 2 forms, along with the field notes, will be sent to Paula Schwanenflugel at UGA.**

Field notes overview

Observers will be present for a 90-minute time period during the literacy block of instruction. So that reliabilities can be determined later by a fidelity checker, who was not present, observers will take extensive field notes on all teacher activities. If there is an assistant in the room, they will also be observed with equal weight to the teacher on a five-minute on-teacher and five-minute on-teacher assistant rotation. It is important that the notes are descriptive enough so that complex observations such as question types, linguistically complex talk, and extension activities can be fully coded. It is incumbent upon the classroom fidelity observer to provide the necessary details so that items can be properly coded.

Field notes will be taken on numbered/lined paper or using a computer generated line document, so that evidence of elements observed can be provided on the *Checklist*. All observers should have the numbered paper available in case they are unable to use their computer (battery dead, too disruptive to classroom, etc.).

Particular emphasis will be noted on questions, elaborations for children's speech, and vocabulary recasting; these will be quoted to the extent possible. Further, if children are sent to a particular center by the teacher or assistant, what the children are asked to do in that center should be noted. Any mention of vocabulary as a topic should be quoted to the extent possible.

Make note of group size and the length in minutes of any activities observed.

TRAINING FIDELITY CHECKLIST DETAILED DIRECTIONS AND IMPLICATIONS FOR NOTE TAKING

Below you will find a description of each numbered component of the *Checklist*. In this description, we have provided a description and brief rationale for the intervention component so that ambiguous cases can be better decided. Each numbered component on the *Checklist* is scored individually. (i.e., a teacher may need remediation only on intervention component #1 and have no need of remediation on component #'s 2 and 3 even though all three items are part of the Building Bridges component). We have also provided recommendations for how field notes should be taken. Keep this information handy as a reference while you are in the classroom.

Building Bridges—child-centered conversation

The purpose of this component is to promote conversations that encourage children to build good relationships with their teachers through talk. When children recognize that these conversations are on topics of interest to them, they can take risks with talk.

a. Children's interests or topics should be dealt with in these conversations. The topics should be non-didactic. That is, teachers should not be using these conversational groups as a way of re-teaching other topics within the curriculum. If academic topics are brought up, they should be brought up by the children first. These topics should be pursued by teacher only to the extent that children seem interested in knowing more. This should be indicated by children continuing to respond to this topic. They should not be cut off so that the teacher can redirect them to her own topic.

b. Children should be allowed a significant opportunity to talk. Teacher should attempt to get child to talk further when limited responses are made. Teacher may emphasize good turn-taking skills among children.

c. The above characterization should be true for the majority of time spent in any five-minute conversation segment observed.

d. These components must be observed.

Field notes should note whether teachers maintain children's topics. They should indicate attempts by teacher to get child to talk more. They should note attempts to turn these into didactic moments on teacher's part (the latter to be viewed as evidence for need for remediation).

Building Bridges—sufficient conversation duration

The goal of this component is to ensure that sufficient amount of teacher-child conversation occurs throughout the school year to encourage the development of language skills.

a. Small group or individual conversations that are five minutes in duration or longer. Group size should be seven or fewer. Minor interruptions (< 1 min.) by children outside the small group may occur during these sessions, but the teacher must demonstrate a good attempt to hold an extended conversation. Note # of children in the group.

b. These can be carried out at a variety of times such as following small group academic times, small group book readings or extension activities, or they can be scheduled separately.

c. At least one conversation must be observed; the second may be counted based on self-report only if one five-minute conversation was observed during the observation. If these conversations occur outside the observation, the second can be self-reported by teacher as having occurred during play times or recess, during early arrivals or late departures, during family-style meals with teacher, during nap time for non-sleepers, or before or after school or with "early finishers" for other academic times.

d. Ideally, two sessions per observations will be observed. However, if one session is observed then you may accept a self-report of the second.

Field notes should indicate time stamp so that duration can be determined. Count the number of children in the group. If your attention shifts to the other adults, please note if the conversation continues. Each instance of extended conversations should be noted. If there is at least one observed during the 90-minute observation period, teacher can self-report a second one at some specific other point in the day. This should be evident in her lesson plans, tracking tool, or if the teacher is able to describe the conversation with sufficient detail.

Building Bridges—linguistically complex talk

The purpose of this component is to encourage teacher speech in forms known to be related to the development of oral language skills and vocabulary in children. It takes a variety of forms including:

a. Complex talk—

i. Cognitively complex questions that query internal states (What do you think he was feeling? Thinking?), clarifying and hypothesizing (What do you mean? What do you think happened?), and open-ended questions (How, why, what else?)

ii. Linguistic expansions/grammatical recasts which take what the child says and adds the missing grammatical elements to it; or extensions which add elaborative information into the child's sentences. The teacher should be non-corrective in providing these expansions and recasts. The child should not really be aware that he is being corrected. Some examples are:

"My daddy, he go Wal-Mart." →

T: "Your daddy went to Wal-Mart? (grammatical recast)

T: "Has your daddy gone to the Wal-Mart near the mall?" (extension)

b. Vocabulary emphasis—Teacher's language should include vocabulary recasting or rare vocabulary words.

i. Rare vocabulary (words that are used infrequently) and/or vocabulary recasts introduced by the teacher in the conversation.

"There a big car."→

T: "Yes, that's an enormous car." (vocabulary recast); or

T: "Is that a sports utility vehicle?" (rare vocabulary)

c. Each component must be observed at least once in order to check the box as present.

In field notes, observers should write down teacher talk that sounds like a modified repetition. They should write down teacher talk that sounds like a deliberate attempt to use a rare word. They should write down as many teacher questions during these conversations as possible.

CAR Talk—competence questions

The purpose of these questions is to allow children to have a successful experience among the various questions that may be asked by the teacher during book reading. These success experiences are designed to encourage child participation during storybook reading and comprehension.

a. Classify question as competence if children would be expected to know the information being queried. This is often information that is right in the book either just previously stated or in the book's picture. A vocabulary word might be embedded, however.

b. Some examples are: Find the (object); What color is the (object)? Where is the (object)? What shape is the (object)? What is this? Tell me about (object, person). What is (character)'s name? What else do you see? What did you just find out? What are (names) doing? Where is (person) (action)? Who said (phrase)? Who is (name) talking to? When did (person) (action)? And (person) said….

c. Questions should be concrete.

d. Two of this type question must be observed.

For field notes, record which book was used and write down all the questions asked during the book reading. The question type can be identified later. Please note, this is not about pacing of questions or if the teacher provides sufficient time for students to respond. If the teachers are "question machines" that can be noted and addressed in the teacher debriefing session but it is not relevant to the implementation assessment.

CAR Talk—abstract questions

The purpose of these questions is to encourage cognitively and linguistically complex talk which has been shown to encourage the development of vocabulary and comprehension. They will also encourage child participation.

a. These are questions which ask students to summarize, define, explain, judge, compare, contrast, predict, take another point of view, or solve problems. *How* and *why* questions and questions querying about the internal states of characters are always abstract.

b. Some examples are: Why do you think (person) did (action)? What is (character) thinking? Who does this remind you of? What does (word) mean? How are (two objects) the same? What will happen next? How are (two objects) different? How do you think (person) feels about (action)? Where else do (people, animals) live? Why is (person) (action)? What was the story about?

c. Two of this type question must be observed.

For field notes, record which book was used and write down all the questions asked during the book reading. The question type can be identified later. Please note, this is not about pacing of questions or if the teacher provides sufficient time for students to respond. If the teachers are "question machines" that can be noted and addressed in the teacher debriefing session but it is not relevant to the implementation assessment.

CAR Talk—relate questions

The purpose of these questions is to encourage children to relate the contents of books to their own lives and prior knowledge. This has been shown to encourage comprehension.

a. These are questions that ask children to relate the contents of books to their own lives.

b. Some examples are: Have you ever had a really horrible day? What do you do when you have a bad dream? What would you do if you could (action) like Hager? Who has on a blue shirt like Oliver's? Whose birthday was it last week? What does your mom like to cook?

c. Two of this type question must be observed.

For field notes, record which book was used and write down all the questions asked during the book reading. The question type can be identified later. Please note, this is not about pacing of questions or if the teacher provides sufficient time for students to respond. If the teachers are "question machines" that can be noted and addressed in the teacher debriefing session but it is not relevant to the implementation assessment.

New Vehicles—quick definitions

The purpose of these questions is to get children used to listening to and thinking about definitions for words, which is a skill which is under development at this age.

a. Teacher supplied definitions of target words or of any words.

b. This would generally occur early in the unit when words are being introduced.

c. This can be self-reported.

For field notes, note any verbal definitions supplied by teacher for target words either within the initial presentation of the targets or later. Count this as observed if you see the teacher providing any quick definitions–even if this is not of a target word. If it is not the first day of the unit, the observer may not see teacher defining targets, so you can rely on self-report for this.

New Vehicles—sufficient small-group book reading quantity

The purpose of the small group book readings is to allow children to have ample opportunity to learn vocabulary words by listening to books being read because written language emphasizes different vocabulary than spoken language.

a. Two small group book readings should be present in any 90 minute observation.

b. Group size should be seven or fewer.

c. If only one is observed, the second can be self-reported.

For field notes, make note of any time a teacher is reading to a small group or individual. Count and make note of the number of children in the small group. There should be two instances of small group readings per day. Do not count any small grouping where children are reading book to teacher or reading silently to themselves. If there is at least one small group reading observed during the 90-minute observation period, the teacher can self-report a second one at some specific other point in the day. This should be evident in her lesson plans or tracking tool.

New Vehicles—N^3C introduction of vocabulary

The purpose of this practice is to provide children with a chance to exercise a strategy for learning to words that is already likely to be in their repertoire. This should provide an opportunity for children to look and feel smart in the learning of new words.

a. The teacher should present target word picture card in the context of several "known" (non-target) picture cards.

b. The teacher should query each by saying "show me (target)" or "who can show me (target)?" or something similar.

c. If not observed, this can be self-reported.

For field notes, if this is not at the beginning of the vocabulary units, observer will not see it. Observer may rely on teacher self-report for this. Do not count any presentation where teacher merely presents the target picture and names it, although this can be done in conjunction with N^3C introduction.

New Vehicles—presence of vocabulary targets

The purpose of this practice is to communicate explicitly to children that there is a list of words that they should be learning.

a. Teacher posted lists or pictures or props of targets as a group somewhere in the room; OR

b. Teacher mentions that there are vocabulary words.

c. One of these must be observed.

For field notes, postings of target list or pictures, Table of target props should be noted. Direct mention of the topic of vocabulary words should be noted. If the teacher has sent home a parent letter with the weekly target vocabulary targets make note of that.

New Vehicles—extension activity

The purpose of this practice is to allow children to have the opportunity to practice the vocabulary words in some activity. Ideally, this activity would be a somewhat authentic context that will enable the children to learn when, where, how, and why particular words might be used.

a. Children were sent to centers where they were encouraged to process vocabulary words through some activity.

b. This must be observed.

For field notes, write down any statements given by teachers regarding what the purpose of a small group activity is. Write down any statements about a center being for vocabulary. Write down directions given to children regarding a center activity that possibly seems related to a vocabulary theme. Do not observe the group activity if there is no adult present, rather note the directions students were given when sent to the center. Do not evaluate the quality of the activity-which may be addressed in the debriefing session-but focus on the presence of the activity, no matter how boring, etc.

New Vehicles—sufficient small group extension activity quantity

The purpose of this practice is to ensure that children have enough practice using the vocabulary words in an activity.

a. At least two groups of children participating in small group activities is necessary.

b. Activities should target vocabulary.

c. If only one is observed the second can be self-reported.

For field notes, note if there are at least two groups sent to these activities. If there is at least one observed during the 90 minute observation period, teacher can self-report a second one at some specific other point in the day. This should be evident in her lesson plans or tracking tool.

FORM COMPLETION–TRAINING FIDELITY CHECKLIST

Complete as much of the *Training Fidelity Checklist* as possible using your field notes. (Attach notes to the checklist for reference.) There are some components of the observations where teacher self-report is allowable. These instances are marked within the *Fidelity Checklist*. In the case where an intervention activity is not observed during the course of the observation, and the *Fidelity Checklist* for this item allows self-report, the teacher will be asked to provide evidence on how that activity was included in a previous lesson during the after school meeting. If the component must be observed to be counted that is noted in the *Fidelity Checklist*.

Use column 2 of the *Checklist* to document whether or not you observed the intervention component. Some items ask for group size or conversation length. Evidence for each intervention component will be noted on the *Checklist* by noting the line number in the field notes where the item was observed.

Use the third column of the *Checklist* to indicate whether remediation may be required (by circling the appropriate designation) and to write any comments, notes, or suggestions that you want to share with the teacher.

As you review the *Checklist*, please remember that each numbered item on the *Checklist* is scored independently—that is, each item stands on its own. For example, although the Building Bridges component consists of three separate practices (child-centered conversation, sufficient quantity, and linguistically-complex talk), a teacher may only need remediation on the item related to linguistically complex talk not the entire Building Bridges component.

A teacher must receive all positive marks (Yes, observed) for each numbered item on the *Checklist* to be considered not to need remediation. For each item that the teacher does not show evidence in implementation (Not observed) remediation will be provided. In order to preserve reliability, the observer should always remember to keep in mind that the presence of a component is what is scored—rather than judgments about the quality of the activity. Each numbered component on the *Checklist* is scored individually (i.e., a teacher may need remediation only on intervention component #1 and have no need of training on component #2 and #3 even though all three items are part of the *Building Bridges* component).

Teacher: _____ School: _____

Observer: _____ Date:_____ Time Observed:_____

(Line to be completed by data manager) Teacher ID#_____ School ID #_____

Training Fidelity Checklist

Intervention Component	Evidence (Insert line numbers from observation field notes.)	Remediation Required? (circle) Notes for debriefing
1. *Building Bridges*—child-centered conversation (p. 10): ☐ Teacher[a] engaged children in conversations centered on children's interests and topics; ☐ Allowed children significant opportunity to talk. *Note: Both components must observed*		Yes/No
2. B*uilding Bridges*—sufficient conversation duration: ☐ Teacher engaged in at least *two* small-group[b] conversations that lasted at least 5 minutes. (Fill in # of min. for each conversation observed.) *Note: If only one observed, second can be self-reported in interview*	1st group # of children in group ____ Conversation length in minutes ____ 2nd group: # of children in group ____ Conversation length in minutes ____	Yes/No
3. *Building Bridges*—linguistically complex talk (p. 11): ☐ Complex talk—Teacher's conversations with children included questions that encouraged interpreting, hypothesizing, clarifying, open-ended questions (how, why, what else); and/or teacher carried out linguistic expansions and extensions. ☐ Vocabulary emphasis— Teacher's language included vocabulary recasting or rare vocabulary words introduced as a natural part of conversation. *Note: Both aspects must be observed*		Yes/No
4. *CAR Talk*—competence questions (p. 16): ☐ When reading a book to children, teacher asked *two* questions where children could demonstrate competence. (Ex. Who said…? Where is…?) *Note: Must be observed*	1st question: 2nd question:	Yes/No
5. *CAR Talk*—abstract questions (p. 16): ☐ When reading a book to children, teacher asked *two* questions where children could demonstrate ability to think abstractly. (Ex. Why do you think…? What will happen…?) Fill in blanks with questions asked. *Note: Must be observed*	1st question: 2nd question:	Yes/No

Teacher: _____ School: _____

Observer: _____ Date:_____ Time Observed:_____

(Line to be completed by data manager) Teacher ID#_____ School ID #_____

6. *CAR Talk*—relate questions (p. 16): ☐ When reading a book aloud, teacher asked *two* questions where children could relate book to their lives. Fill in blanks with questions asked. *Note: Must be observed*	1st question: 2nd question:	Yes/No
7. *New Vehicles*—Quick definitions: ☐ The teacher supplied the definitions of vocabulary words or other words used. *Note: If not observed, can be self-reported in interview*		Yes/No
8. *New Vehicles*—Sufficient small group book[b] reading quantity (pp. 14 & 17): ☐ Teacher carried out two small-group interactive readings of books. *Note: If only one small group reading is observed, second can be self-reported in interview*	1st reading: # of children in group ____ 2nd reading: # of children in group ____	Yes/No
9. *New Vehicles—N³C* introduction of vocabulary (p. 20): ☐ Teacher engaged children in a novel-name-nameless category activity. *Note: If not observed, can be self-reported in interview*		Yes/No
10. *New Vehicles*—Presence of vocabulary targets (pp. 19-20): ☐ This teacher utilized a targeted list of vocabulary words for the unit of instruction; or ☐ Teacher mentions the topic of vocabulary words *Note: Must be observed*		Yes/No
11. *New Vehicles*—Extension activity (p. 21): ☐ Children were sent to centers which encouraged use of vocabulary words in activities. *Note: Must be observed*		Yes/No
12. New Vehicles—Sufficient small group extension activity quantity (p. 21): ☐ At least *two* small groups[b] were sent to centers where they could practice vocabulary. *Note: If only one is observed, second can be self-reported in interview*	1st activity # of children in group ____ 2nd activity # of children in group ___	Yes/No

[a] *Teacher in all instances means either lead or assistant teacher. Analysis is at classroom level, so any instance by either teacher counts.*

[b] *Small group* in all instances means group size of 7 or fewer. Page numbers noted refer to *K-PAVEd for Success Teacher's Guide*.

TEACHER OBSERVATION FOLLOW-UP MEETING PROTOCOL—DETAILED DESCRIPTION

Fidelity observers will meet individually with each teacher observed immediately after school (for approximately 20-30 minutes per teacher) on the day of observation. Teacher assistants are not required to be in the follow-up debriefing but are welcome to participate if they wish.

Preparing for the follow-up/debriefing session

The second form, the *Teacher Observation Follow-Up Meeting Protocol* is used during the after school meeting with the teacher. Prior to the teacher debriefing, observers should carefully review the classroom observation *Checklist* and field notes taken during the observation, and complete the *Follow-Up Meeting Protocol* form.

The purpose of this debriefing is to provide the teachers with feedback on the intervention delivery in their classroom. Our goal is to use timely and elaborative feedback to reinforce good performance, shape and strengthen more tenuous performance, and to investigate, explain, and possibly remediate poor or absent behavior.

Suggestions for conducting the follow-up/debriefing session

As you complete the *Protocol*, remember, the more specific feedback that teachers receive the better opportunity they have to change or correct their teaching. Try to provide teachers with enough detail so that they may either continue their current teaching behaviors or work to improve those behaviors. (The following directions are numbered the same as the *Protocol* form for ease of use.)

I. Thank You–Begin the debriefing session by thanking the teacher for their participation in the study. The *Protocol* contains a suggested paragraph (Item I) you can read at this time. It may be helpful to explain to the teacher the process you used to collect information and how the *Checklist* is used to assess the teachers' effectiveness in implementing the K-PAVE program.

II. Follow-Up/Debriefing–Item II of the *Protocol* has six items to discuss during this meeting. Use the *Protocol* in conjunction with the *Checklist* to guide the discussion on these six items.

1. Begin by noticing a positive that is not be related to the intervention. (e.g., "I really liked the way you had children's artwork exhibited" or "You did a great job of holding the children's interest during circle time this morning).

2. Focus on an element that was particularly strong. Provide detail on how this aspect was done especially well.

3. Go through those areas that appeared to be strengths for the teacher. Hopefully, this will encourage him/her to continue to use practices that are appropriate and effective. Move through each item that was marked as observed (No – Remediation

required) on the *Checklist* and briefly discuss those items with the teacher offering as many positive illustrations as possible.

4. Go through the aspects of the program that they got almost right and re-explain where needed. This would pertain to items that were marked as observed on the *Checklist* but needed some improvement or could have been done with a higher level of quality. (e.g., "You asked the right type of questions during your book reading, but it might have been better if you had allowed the children a little more time to answer your questions.") Try to include specific examples of where and how the teacher's practices could be improved. Try to determine if you are dealing with a practice for which the teacher understands the underlying concepts and principles but has not yet had sufficient practice, versus a situation in which the teacher has not mastered the prerequisite skills necessary to successful performance of the desired behavior. This distinction can guide the nature of the feedback and/or remediation you offer, namely additional practice versus a more basic revisitation of the foundational skills and concepts (Item 6 below).

5. Ask about the aspects you didn't see that allow for teacher self-report. The items that allow for teacher self-report are clearly marked in the *Checklist*. If they self-report any items, be sure to check if remediation is needed for that item.

6. The observer should focus on components of the intervention that were marked as needing remediation. Remediation simply means that there were some items the teacher could use a second explanation on how to make this part of the intervention work. Below are some specific suggestions that you can use to help teachers better understand the skills that were taught during the training:

- Provide feedback on what particular practice might have been missing from the teacher's performance or what practice might have been performed incorrectly.
- Ask the teacher to open to the section in question within the Teacher's Guide.
- Together, you and the teacher move through the concepts and ideas that were presented during the training.
- Ask the teacher to indicate where difficulty was encountered or to discuss which ideas were unclear or which presented the greatest challenge.
- If the difficulty appears to be a failure to understand or comprehend a particular concept (e.g., child-centered conversation), then the observer may need to provide further elaboration on the content or provide additional examples to illustrate the concept.
- If the difficulty appears to be more of a problem with procedures (e.g., presenting N3C pictures), then it may be helpful for the observer to demonstrate the practice and then allow the teacher practice.

III. Implementation Challenges–There are six questions that you should ask the teachers during the after-school debriefing session. Write the teacher responses in the space provided. The *Protocol* form and the *Checklist* will both be sent to Paula Schwanenflugel at the University of Georgia in order to help guide the telephone follow-ups with teachers.

IV. Reminder of Follow-ups and Contact Information (Both in person and via telephone with the University of Georgia.)—Ask the questions and obtain contact information for this item on the *Protocol.* Give the teacher your contact information in case he/she would like to contact you at a later date.

Teacher: _____ School _____

Observer: _____ Date: _____

TEACHER OBSERVATION

FOLLOW-UP MEETING PROTOCOL

Thank You (please read this or rephrase as you wish)

"Thanks for participating in this study and attending the training. We know that there are many things that teachers are asked to do surrounding literacy. Without teachers willing to take what they learned in training and try it out in their classrooms, we couldn't make progress in designing effective programs for young children. We know it is hard to implement this program just a day or so after training. You probably haven't had the chance to fully digest the teachers' manual and thoroughly examine the vocabulary units yet. I am here to help get you started with both tasks. I am not here to provide an in-depth critique of your teaching, but rather to get a sense of what I can do to support your work in implementation of the K-PAVE materials. I also want you to know that my observation and this conversation will not be relayed to any of your school or district staff."

II. Training Follow-up

1. Notice something in his/her teaching *outside of the intervention*. "First of all, I really loved _____ (e.g., her nice manner with the children, her good classroom management, her effective organization of the room). (elaborate)

2. Focus on an element of the *intervention* that was particularly strong. "With regards to the program, I thought the way you carried out _____ (some aspect of the program) was just terrific." (elaborate)

3. Go through the positive aspects of the *Fidelity Checklist*. "Even though it has only been hours since you were introduced to this new curriculum, I notice that you did a great job in _____ (some practice)"

4. Go through the aspects of the program that they got almost right and re-explain. "I saw you _____ (some practice). This was pretty close to what we had in mind, but_____ (Describe fully correct practice)."

5. Ask about the aspects you didn't see. "We were looking for _____ (some aspect of program), but we didn't see it. Did you carry this out at some other time when we weren't there? (yes or no). What did you do? _____
_____
_____ (teacher response).

For items marked as Remediation Needed (See **Follow-up/Debriefing and Remediation – Detailed description II.7** above for suggestions on how to address these items.)

Re-explain any marked items.____ (check, if second explanation was all that was needed)

If they admit to not carrying out a particular practice, re-explain the rationale for the practice and go over part of the teacher manual that indicates what to do _____ (check, if appropriate).

III. Implementation challenges:

(a) Are there aspects of this program you had difficulty implementing?

(b) Are there aspects of this program in which you don't see a purpose?

(c) Are there aspects of the program you really like?

(d) Are there aspects of the program you really can't envision using? If so, what might make it easier for you to work on including those practices?

(e) Are there any aspects of the program you think are developmentally inappropriate for children of this age?

Do you have any suggestions for how we can make the program easier for you to use?

IV. Reminder of follow-ups:

1. We will be coming back to visit again within the next few weeks. Is that okay with you? _____ (yes/no; teacher response). Would you be willing to carry out all the aspects of the program at that time so we can see that you know what to do?" _____ (yes/no; teacher response).

2. For all teachers, "We will be carrying out some follow-ups by phone where you can ask us questions, listen to experiences of the other teachers, and so forth. These will occur about once a month after school. We handed you a personalized schedule yesterday that let you know when your phone calls have been scheduled. Can we have some basic contact information such as your phone number and email?"

Contact Information for Teacher: _____

Phone number: _____

Email address:

_____

Do you monitor this email fairly often? Yes/No

FREQUENTLY ASKED QUESTIONS

(As we get more questions we will continue to add them and distribute to all.)

1. What do I do if there is no teacher assistant in the classroom or the teacher assistant didn't show up for training?

    a. Rather than rotating 5 minutes on teacher and 5 minutes on teacher assistant you will observe the teacher for the entire 90 minutes. Be sure to clearly mark on your forms that only the teacher was observed.

2. What do you tell the principal if they want to know details about your observation?

    a. Due to the nature of the research we are not able to report out on any particular classrooms or teachers. You should also let the teachers know that you will not be discussing your observation or your follow-up discussion with anyone else at their school or district.

3. What do I do if the 90-minute period is interrupted by a special class (i.e., gym, music, art) or by any other interruption (fire drill, etc.)?

    a. Go back and continue you observation so that you have a full 90 minutes in each classroom. Make sure that your field notes state that there was an interruption.

4. What do I do if the teacher can't stay after school for the follow-up meeting?

    a. Hopefully this won't happen, but if it does see if you can reschedule it another time during the same day. Perhaps the principal or someone else could watch the class while you conduct this follow-up as it only takes a few minutes to conduct.

APPENDIX I. DATA COLLECTION PROCEDURES

PROTOCOL FOR CHILD ASSESSMENTS: QUICK REFERENCE

Introducing yourself to the student

Required:

Hi, (student's first name), my name is (your name).

I'd like to talk with you today and show you some pictures and ask you some questions and listen to some stories together. I'll be talking with some other kids today too. And I have some stickers to give you when we're finished.

Required:

When I ask you questions, I am going to write down your answers, but I'm not going to tell your teacher or the other kids what you say. I just want to learn more about what kids your age know.

Would it be okay for you to come with me [to the library, my room, the place designated by the school]? I'd like to learn more about what you know about words and stories. You don't have to go with me if you don't want to. You can let me know anytime when you want to go back to your room.

If child says "YES⇨ Proceed to the testing location.

If child says "NO"⇨ Say, "Okay, I'll check back with you later."

****

Would you feel better if [**name of assistant teache**r] walked over with us or would you like to walk over with me yourself?

Build rapport while walking to the assessment location

Choose one or two:

How old are you? What did you do on your last birthday?

Do you have any brothers and sisters? What are their names? How old are they?

What kinds of things do you like to do when you're not in kindergarten/school?

Do you have any pets? What pets do you have?

I see you have Hulk/ Spiderman/Wall-E/dinosaurs… etc. on your T-shirt/ shoes… etc.. Do you like Hulk/ Spiderman/Wall-E/dinosaurs?

I see you have ribbons in your hair today/special shoelaces/etc.

Explain the assessment processI'll need you to sit up straight and tall in that chair. I'm going to show you some pictures and ask you some questions. Listen carefully and give your best answer. Sometimes the questions might seem hard, but that's okay, some of these

questions are for older kids. If you are not sure what to answer it's okay to take your best guess?

(Continue with the following.) Also, I have to follow some rules, too. One of the rules says that I'm not allowed to tell you whether you're right or wrong. Do you have any questions before we get started? Do you need to go to the bathroom first? Ok, if you need to use the restroom, just let me know. Let's get started.

- WJ-III Passage Comprehension
- WJ-III Academic Knowledge
- EVT–2
- KTEA Listening Comprehension
- If applicable, Language Elicitation Task

Protocol for Classroom ObservationsWhen you arrive at the classroom, you will begin by talking with the teacher (see "General Guidelines for Observing in Classroom" tab). At this time, you will remind the teacher and assistant teacher that they have each agreed to complete a short, self-administered survey. The teacher and assistant teacher will each be given a survey to complete on their own – the Teacher Survey and the Assistant Teacher Survey, respectively (see "Teacher Surveys" tab). You will collect the surveys from the teacher and assistant teacher when the observation is complete, and you will take the surveys with you when you leave the school.

In addition to giving the teacher the Teacher Survey, you will also remind the teacher that we will be recording her speech during the observation (see "Teacher Speech Sample" tab). You should show the teacher the digital recorder, explaining that the lanyard enables the recorder to be worn around the neck and that the clip secures the recorder so that it does not swing. You can work with the teacher to adjust the length of the lanyard. Be sure to turn the recorder on before giving it to the teacher to wear.

Turn the recorder on by pushing the red "REC" (record) button on your digital recorder. You will begin by providing the identifying information for that session. Speak clearly as you give the following information:

- Teacher's name
- Teacher ID number
- Date of observation
- "Conducted by**,** (your name and your observer ID number)"

Repeat this information a second time to ensure that it will be clearly understood by other research staff listening to your recording. DO NOT STOP THE RECORDER, BUT LET IT CONTINUE TO RECORD. Then ask the teacher to wear the recording device. Ask the teacher to slip the lanyard holding the recorder around his or her neck and to clip the device to a piece of clothing so that it does not bounce around during instruction.

Once you have given the teacher and assistant teacher the surveys and given the teacher the digital recorder to wear, get prepared to start your observation session.

Familiarize yourself with the classroom layout, and identify areas that will allow you to easily see and hear what is going on in the class while causing minimal distraction to the teacher and students. Make sure that your recording sheets are organized in a manner that it will be easy for you to change between them quickly.

Since you will be standing and moving around the classroom during the observation session, it is useful to have a clipboard for writing. It is easy to shift between instruments efficiently using the following order of materials on the clipboard from top to bottom: CLASS booklet, folded to the current observation sheet; the Vocabulary Record sheet; the RAP-K; and finally, the laminated CLASS dimensions overview fold-out. Underneath everything you should have your CLASS manual with you; you may need to consult the manual when scoring CLASS cycles. The CLASS form sits on top since most of the observation session will be recorded on it. The Vocabulary Record sheet is next, so when vocabulary instruction occurs it is easy to flip to the sheet and record. When it is time to use the RAP-K, it can be pulled out and placed on top while coding and then returned in the stack once completed. The dimensions overview stays on the bottom of the stack, since it will only be pulled out to use while you are scoring the CLASS.

If you are scheduled to start your observation at a specific time, begin at that time. Most observations will be scheduled to begin at the start of the school day as students arrive. (You should always arrive at the classroom well in advance of the start of your observation.) You will most likely begin observing using the CLASS – not the RAP-K. If beginning your observation at the start of the school day, wait until at least four students arrive and then start the first cycle of the CLASS. Even if formal instruction is not taking place yet, remember that the CLASS dimensions reference many other aspects of the classroom experience that would be a part of other morning activities.

APPENDIX J. DATA QUALITY ASSURANCE PROCEDURES

Student assessments and classroom observation data underwent a thorough quality assurance protocol. Data quality monitors checked each individual completed instrument, using protocols developed by data collection trainers to ensure that forms were completed correctly. The quality control process eliminated any out-of-range values for student and classroom outcomes by examining the raw data and correcting scoring errors.

When the quality control monitor discovered more serious errors that could not be easily corrected, such as a child assessment for which a basal was not established, the completed score sheets were sent to data collection trainers for further review. Trainers determined whether an estimate of the total raw score could be made based on the completed items or whether a raw score could not be imputed. If possible, a raw score was imputed based on the completed test items. All tests with administration errors for which raw scores were imputed based on the completed items were flagged, and, as a sensitivity analysis, impacts were estimated both with and without tests with imputed scores.

For classroom measures (Classroom Assessment Scoring System [CLASS], Read Aloud Profile–Kindergarten [RAP–K], and Vocabulary Record), no individual items used to create the total scores were missing. Either the instrument was completed as part of the classroom observation or the entire instrument was missing. (Specifically, if the observation was conducted, there were three to five CLASS cycles, and there was one completed RAP-K. If a book read-aloud did not occur during the observation, the classroom had valid scores on the RAP-K of 0 for comprehension support, 0 for open-ended questions, and 0 words introduced.)

Standardized student assessment scores were calculated from raw scores electronically to eliminate computation errors. For the two subtests of the Woodcock-Johnson III/Normative Update, academic knowledge and passage comprehension (which will be examined in an exploratory analysis presented in a later report), raw scores were converted to W-scores, which are item response theory–based scale scores, using the Woodcock-Johnson Compuscore Program. Students' raw scores, dates of birth, dates of testing, and gender were entered into the program, which generated the W-scores. To check the accuracy of the data entry, W-scores were compared with raw scores from the original data file, and any inconsistencies were flagged, double-check, and rectified, if necessary. Classroom instruction variables, which were created by summing and averaging observer ratings or tallies, were all cross-checked to ensure that variables were created correctly.

Once analytic variables were created, descriptive analyses of all outcome and covariate measures were conducted. The distribution of each measure was examined for any out-of-range values and outliers. Although missing values were present (their handling is described in Chapter 2 and Appendix M), out-of-range values were not observed for any classroom instruction pretest or posttest variables or for any student covariates, teacher covariates, or school covariates. For student outcomes, seven students had a raw score of 0 either at pretest or posttest on at least one assessment of expressive vocabulary or academic knowledge.[65] Hard copy score forms and assessor notes were examined to confirm that the 0

[65] Raw scores of 0 for listening comprehension were present and were not unusual.

raw scores were accurate for each of these students. Because there were no notes from assessors indicating that the student refused to complete the assessment, the 0 scores were retained. A sensitivity analysis was conducted to examine the influence of students who provided no responses when tested (see Appendix N).

APPENDIX K. MODEL SPECIFICATIONS

THREE-LEVEL MODEL USED TO ESTIMATE IMPACTS ON KINDERGARTEN STUDENTS (FOR SINGLE OUTCOME MEASURES)

To model the impact of the K-PAVE intervention on kindergarten students, a hierarchical linear model was estimated. This model provides an estimate of the average impact of the intervention on students across all schools at a given time point (for example, at the end of the kindergarten year) as well as an estimate of this impact's standard error. In the evaluation data, students are nested within classrooms, and classrooms are nested within schools. Therefore, a three-level hierarchical linear model was specified, with students nested within classrooms within schools. The multilevel modeling also parses the variance among students, classrooms, and schools to produce both more precise point estimates of intervention impact and more accurate standard errors (Raudenbush and Bryk 2002).

The model used to test impacts on students is written in hierarchical form, as shown below. Student and school covariates are defined in Appendix O.

Sampling weights were used to adjust for the loss of one school that dropped out of the study. The weighting results in estimates that represent the sample of 65 schools that were randomly assigned rather than the analytic sample of 64 schools—missing the one school that dropped out. Sampling weights were constructed for the block of intervention schools without either Reading First or a Mississippi reading initiative (that is, the school had a local reading program or no reading program). The 20 schools remaining in this block after the loss of one school were weighted 1.05, so that estimates would represent the full sample of schools that were randomly assigned. All schools in the other blocks were assigned a weight of 1.0.

The level 1, or student-level model is:

(K1)

$$
\begin{aligned}
Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(pre_{ijk} - \overline{pre}) + \beta_{2jk}(female_{ijk} - \overline{female}) \\
&+ \beta_{3jk}(StudentIEP_{ijk} - \overline{StudentIEP}) + \beta_{4jk}(FreeReducedLunch_{ijk} - \overline{FreeReducedLunch}) \\
&+ \beta_{5jk}(AfricanAmerican_{ijk} - \overline{AfricanAmerican}) + \varepsilon_{ijk}
\end{aligned}
$$

where Y_{ijk} is an outcome measure (for example, Expressive Vocabulary Test–2 score) of the i^{th} student in the j^{th} classroom in the k^{th} school; pre_{ijk} is the baseline version of the outcome measure for the i^{th} student in the j^{th} classroom in the k^{th} school (centered at the grand mean, $\overline{pre}$); $female_{ijk}$ is a dummy variable taking the value 1 if the i^{th} student in the j^{th} classroom in the k^{th} school is female and 0 otherwise (centered at the grand mean, $\overline{female}$); $StudentIEP_{ijk}$ is a dummy variable taking the value 1 if the i^{th} student in the j^{th} classroom in the k^{th} school has an receives special education services (that is, has an Individualized Education Plan) and 0 otherwise (centered at the grand mean, $\overline{StudentIEP}$); $FreeReducedLunch_{ijk}$ is a dummy variable taking the value 1 if the i^{th} student in the j^{th}

classroom in the k^{th} school is eligible for free or reduced-price meals and 0 otherwise (centered at the grand mean, $\overline{FreeReducedLunch}$); $AfricanAmerican_{ijk}$ is a dummy variable taking the value 1 if the i^{th} student in the j^{th} classroom in the k^{th} school is African American and 0 otherwise (centered at the grand mean, $\overline{AfricanAmerican}$); β_{0jk} is the covariate adjusted mean value of the outcome measure for classroom j in the k^{th} school, β_{1jk} through β_{5jk} are regression coefficients indicating the effect of each student-level covariate on the outcome measure Y_{ijk} ; and ε_{ijk} is the student-level residual or error term of the i^{th} student in the j^{th} classroom in the k^{th} school (the assumed distribution of these residuals is normal, with mean 0 and variance ϕ^2).

The level 2 or classroom-level model is:

(K2)

$$\beta_{0jk} = \pi_{00k} + r_{jk}$$

where π_{00k} is the covariate adjusted mean value of the outcome measure for school k, and r_{jk} is the error term of the j^{th} classroom in the k^{th} school (the assumed distribution of these residuals is normal, with mean 0 and variance σ^2).

The level 3 or school-level model is:

(K3)

$$\begin{aligned}\pi_{00k} = {} & \gamma_{000} + \gamma_{001}(T_k) + \gamma_{002}(ReadingFirst_k - \overline{ReadingFirst}) \\ & + \gamma_{003}(MSStateInit_k - \overline{MSStateInit}) + \gamma_{004}(AchLvlIndex_k - \overline{AchLvlIndex}) \\ & + \gamma_{005}(PctAfrAm_k - \overline{PctAfrAm}) + \gamma_{006}(PctFreeLunch_k - \overline{PctFreeLunch}) \\ & + \gamma_{007}(SmallTown_k - \overline{SmallTown}) + \gamma_{008}(LargeTown_k - \overline{LargeTown}) \\ & + \gamma_{009}(Delta_k - \overline{Delta}) + \upsilon_k\end{aligned}$$

where γ_{000} is the covariate adjusted mean value of the outcome measure across control schools; γ_{001} is the mean difference in the covariate-adjusted outcome between treatment and control schools (main effect of treatment); T_k is treatment status dummy variable taking the value 1 for a school assigned to the K-PAVE treatment and 0 for a school assigned to the control group; $ReadingFirst_k$ is a dummy variable taking the value 1 if the k^{th} school participated in the Reading First program in the 2008/09 school year and 0 otherwise (centered at the grand mean $\overline{ReadingFirst}$); $MSStateInit_k$ is a dummy variable taking the value 1 if the k^{th} school has a Mississippi reading initiative and 0 otherwise (centered at the grand mean, $\overline{MSStateInit}$); $AchLvlIndex_k$ is the Achievement Level Index[66] for the k^{th}

[66] The Achievement Level Index is created by the Mississippi Department of Education for each school in the state, based on student performance on the Mississippi state accountability test (that is, Mississippi Curriculum

school (centered at the grand mean, $\overline{AchLvlIndex}$); $PctAfrAm_k$ is the percentage of students in the k^{th} school that are African American (centered at the grand mean, $\overline{PctAfrAm}$); $PctFreeLunch_k$ is the percentage of students in the k^{th} school eligible for free or reduced-price meals (centered at the grand mean, $\overline{PctFreeLunch}$); $SmallTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a small town and 0 otherwise (centered at the grand mean, $\overline{SmallTown}$); $LargeTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a large town or on the fringe of a city and 0 otherwise (centered at the grand mean, $\overline{LargeTown}$)[67]; $Delta_k$ is a dummy variable taking the value 1 if the k^{th} school is in the Delta and 0 otherwise (centered at the grand mean, $\overline{Delta}$); γ_{002} through γ_{009} are regression coefficients indicating the effect of each school-level covariate on the covariate-adjusted mean value of the outcome measure; and υ_k is the error term for the k^{th} school (the distribution is assumed to be normal, with mean 0 and variance τ^2).

The parameter γ001 indicates the impact of the K-PAVE treatment on the specified student outcome. A t-test was conducted to test the null hypothesis that the treatment effect is 0 using a .05 level criterion to reject the null hypothesis. A positive and statistically significant estimate of γ001 indicates that there is compelling scientific evidence (at the 5 percent level) that the K-PAVE intervention improves student vocabulary (or comprehension) outcomes. The magnitude of γ001 estimates the magnitude of the impact—that is, school participation in K-PAVE is estimated to have, on average, a γ001 point effect on the scores of the students in participating schools.

The standardized effect size is calculated by dividing the estimated impact from the model by the standard deviation of the outcome variable, Y_{ijk}, in the control group, as recommended in Burghardt, Deke, Kisker, Puma, & Schochet (2009) because the intervention might affect the standard deviation in the treatment group. Letting S_c be standard deviation of the outcome measure in the control group means that the effect size is $\frac{\hat{\gamma}_{001}}{Sc}$.

MODELS SPECIFICATION FOR GLOBAL *F*-TEST OF JOINT IMPACT ON MULTIPLE STUDENT OUTCOMES WITHIN A DOMAIN

The secondary confirmatory analysis examines two student outcomes—academic knowledge and listening comprehension—in a single domain (that is, vocabulary-related student outcomes). Before analyzing either measure individually, a joint test of the null

---

Test), which is administered to all students in grade 3 or higher. Scores at all grade levels and for all subject areas are included in the index. The percentage of students in the school who score basic or higher and the percentage of students in the school scoring proficient or higher are used to create an Achievement Level Index score ranging from 100 to 600, with scores in the 100 range corresponding to a school performance level of "low" and scores in the 500 range corresponding to a school performance level of "superior."

[67] Locale is represented by a series of three dummy variables, indicating whether the school is in a small town (*SMALLTOWN*), is in a large town or on the fringe of a midsize city (*LARGETOWN*), or is in a rural area (*RURAL*). The reference category in the model is *RURAL*.

hypothesis that the impact of K-PAVE on both outcomes is 0 was conducted. To do this, the following ordinary least squares regression model for both outcomes within the domain was simultaneously estimated. Values for all student-level variables (that is, posttest, pretest, and covariates) were averaged for each school, and an ordinary least squares regression model was estimated using school mean values.[68] School means were weighted based on the number of students sampled in each school. In addition, weights were included to adjust for the loss of the school that dropped out of the study (as described in the previous section).

(K4)

$$\overline{Y_k} = \beta_0 + \beta_1 T_k + \beta_2 \overline{PRE_k} + \beta_3 \overline{female_k} + \beta_4 \overline{StudentIEP_k} + \beta_5 \overline{FreeReducedLunch_k} + \beta_6 \overline{AfricanAmerican_k} + \beta_7 ReadingFirst_k + \beta_8 MSStateInit_k + \beta_9 AchLvlIndex_k + \beta_{10} PctAfrAm_k + \beta_{11} PctFreeLunch_k + \beta_{12} SmallTown_k + \beta_{13} LargeTown_k + \beta_{14} Delta_k + \varepsilon_k$$

where $\overline{Y_k}$ is an average value of a student outcome measure (for example, an academic knowledge test score) for the k^{th} school; T_k is a dummy variable for treatment status taking the value 1 for a school assigned to the K-PAVE treatment and 0 for a school assigned to the control group; $\overline{PRE_k}$ is the average baseline achievement test score for the k^{th} school; $\overline{female_k}$ is the proportion of students sampled from the k^{th} school that are female; $\overline{StudentIEP_k}$ is the proportion of students sampled from the k^{th} school that receive special education services; $\overline{FreeReducedLunch_k}$ is the proportion of students sampled from the k^{th} school who are eligible for free or reduced-price meals; $\overline{AfricanAmerican_k}$ is the proportion of students sampled from the k^{th} school who are African American; $ReadingFirst_k$ is a dummy variable taking the value 1 if the k^{th} school participated in the Reading First program in the 2008/09 school year and 0 otherwise; $MSStateInit_k$ is a dummy variable taking the value 1 if the k^{th} school has a Mississippi reading initiative and 0 otherwise; $AchLvlIndex_k$ is the Achievement Level Index for the k^{th} school; $PctAfrAm_k$ is the percentage of all students in the k^{th} school that are African American; $PctFreeLunch_k$ is the percentage of all students in the k^{th} school that are eligible for free or reduced-price meals; $SmallTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a small town and 0 otherwise; $LargeTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a large town or on the fringe of a city and 0 otherwise[69]; $Delta_k$ is a dummy variable taking the value 1 if the k^{th} school is in the Delta and 0 otherwise; β_0 is the covariate adjusted mean value of the outcome measure for control schools whose covariate values are all 0; β_1 is the

[68] Because we could not estimate a series of multilevel models simultaneously, models were estimated using school means for each variable. Using PROC SYSLIN in SAS, we estimated multiple ordinary least squares equations simultaneously and tested the null hypothesis that the impact in all models is 0.

[69] As in the multilevel model, the omitted category for locale is *RURAL*, which indicates that the school is located in a rural area.

average treatment effect; β_2 through β_{14} are regression coefficients indicating the effect of each covariate on the outcome; and ε_k is the error term for the k^{th} school (the assumed distribution of these residuals is normal, with mean 0 and variance σ^2).

This model was estimated simultaneously for both outcomes in the domain—academic knowledge and listening comprehension—using two outcome equations. Correlations among error terms in the two equations were assumed to be 0. The equations were estimated simultaneously in order to be able to impose the null hypothesis that the treatment effect in both equations is 0 and to be able to conduct a joint *F*-test of the hypothesis, which is stated as

(K5)

$$H_0 = \beta_{1(ACADEMIC\_KNOWLEDGE)} = 0 \text{ and } \beta_{1(LISTENING\_COMPREHENSION)} = 0$$

If the *p*-value for this hypothesis test was less than .05, then the null hypothesis was rejected and a multilevel model (as described above) was estimated to test the impact of K-PAVE on each of the individual outcomes. If the *p*-value for this hypothesis test was greater than .05, then the null hypothesis was not rejected and the impact of K-PAVE on the individual vocabulary-related student outcomes was not examined.

TWO-LEVEL MODEL USED TO ESTIMATE IMPACTS ON CLASSROOM INSTRUCTION (FOR SINGLE OUTCOME MEASURES)

The impact of the K-PAVE intervention on classroom instructional practices, controlling for teacher and school characteristics, was estimated using a multilevel model, to account for the clustering of classrooms within schools. The model includes a classroom level (level 1) and a school level (level 2). Because of the limited degrees of freedom at level 1 (due to sampling only two classrooms per school), teacher characteristics were controlled for at the school level. For each teacher characteristic, the average value for the school is calculated.

The multilevel model used to test impacts on classroom instruction took the following form. Teacher and school characteristics used as covariates are defined in Appendix O.

As with the models estimating impacts on students, sampling weights were used to adjust for the loss of one school that dropped out of the study (see above). The weighting results in estimates that represent the sample of 65 schools that were randomly assigned rather than the analytic sample of 64 schools—missing one school that dropped out.

Level 1, classroom-level model:

(K6)

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

where Y_{ij} is an outcome measure for the i^{th} classroom in the j^{th} school, β_{0j} is the mean value of the outcome Y for school j, and ε_{ij} is the residual for the i^{th} classroom 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:

(K7)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(Tch\_\text{Female}_j - \overline{\text{Tch\_Female}}) + \gamma_{03}(Tch\_AfrAm_j - \overline{Tch\_AfrAm})$$
$$+ \gamma_{04}(College_j - \overline{College}) + \gamma_{05}(GradDegree_j - \overline{GradDegree})$$
$$+ \gamma_{06}(CertEC_j - \overline{CertEC}) + \gamma_{07}(Cert\,\text{Re}ad_j - \overline{Cert\,\text{Re}ad})$$
$$+ \gamma_{08}(YrsTch_j - \overline{YrsTch}) + \gamma_{09}(\text{YrsTchKg}_j - \overline{\text{YrsTchKg}})$$
$$+ \gamma_{10}(ReadingFirst_j - \overline{ReadingFirst}) + \gamma_{11}(MSStateInit_j - \overline{MSStateInit})$$
$$+ \gamma_{12}(AchLvlIndex_j - \overline{AchLvlIndex}) + \gamma_{13}(PctAfrAm_j - \overline{PctAfrAm})$$
$$+ \gamma_{14}(PctFreeLunch_j - \overline{PctFreeLunch}) + \gamma_{15}(SmallTown_j - \overline{SmallTown})$$
$$+ \gamma_{16}(LargeTown_j - \overline{LargeTown}) + \gamma_{17}(Delta_j - \overline{Delta}) + \upsilon_{0j}$$

where γ_{00} is the grand mean of the outcome measure for control schools; γ_{01} is the average treatment effect on the classroom outcome; T_j is a dummy variable for treatment status taking the value 1 for a school assigned to the K-PAVE treatment and 0 for a school assigned to the control group; γ_{02} through γ_{09} are regression coefficients indicating the effects of teacher characteristics on the outcome, averaged for school *j*; γ_{10} through γ_{17} are regression coefficients indicating the effects of school characteristics on the outcome, for school *j*; $Tch\_Female_j - \overline{Tch\_Female}$ is the proportion of focal teachers[70] in the j^{th} school who are female, centered at the grand mean; $Tch\_AfrAm_j - \overline{Tch\_AfrAm}$ is the proportion of focal teachers in the j^{th} school who are African American, centered at the grand mean; $College_j - \overline{College}$ is the proportion of focal teachers in the j^{th} school whose highest level of education[71] is a bachelor's degree, centered at the grand mean; $GradDegree_j - \overline{GradDegree}$ is the proportion of focal teachers in the j^{th} school who have a graduate degree, centered at the grand mean; $CertEC_j - \overline{CertEC}$ is the proportion of focal teachers in the j^{th} school with a teaching certificate in early childhood, centered at the grand mean; $Cert\text{Re}ad_j - \overline{Cert\text{Re}ad}$ is the proportion of focal teachers in the j^{th} school with a teaching certificate in reading, centered at the grand mean; $YrsTch_j - \overline{YrsTch}$ is the average number of years that focal teachers in the j^{th} school have been teaching children (that is, total

[70] "Focal teachers" refers to the two teachers in the classrooms randomly selected in each school for data collection. All schools in the study had at least two kindergarten classrooms. In schools with only two kindergarten classrooms, both classrooms were selected for data collection with certainty. In schools with more than two kindergarten classrooms, two classrooms were randomly selected for data collection. In this case, teacher characteristics are averaged for the two teachers from whom data were collected. Data on other kindergarten teachers in the school were not collected and thus are not included.

[71] Highest level of education is represented by a series of dummy variables: *College* indicates that teachers' highest level of education is a bachelor's degree, and *GradDegree* indicates that teachers' highest level of education is a graduate degree. The reference category is teachers with some graduate coursework but no degree obtained. In the sample, 38% of teachers have a bachelor's degree, 25% have some graduate coursework, and 36% have a graduate degree (see table 2.10 in chapter 2.)

years of experience teaching), centered at the grand mean; $YrsTchKg_j - \overline{YrsTchKg}$ is the average of the total number of years that focal teachers in the j^{th} school have been teaching kindergarten, centered at the grand mean; $ReadingFirst_j$ is a dummy variable taking the value 1 if the j^{th} school participated in the Reading First program in the 2008/09 school year and 0 otherwise, centered at the grand mean; $\overline{ReadingFirst}$; $MSStateInit_j$ is a dummy variable taking the value 1 if the j^{th} school has a Mississippi reading initiative and 0 otherwise, centered at the grand mean, $\overline{MSStateInit}$; $AchLvlIndex_j - \overline{AchLvlIndex}$ is the Achievement Level Index for the jth school, centered at the grand mean; $PctAfrAmer_j - \overline{PctAfrAmer}$ is the percentage of students in the j^{th} school who are African American, centered at the grand mean; $PctFreeLunch_j - \overline{PctFreeLunch}$ is the percentage of students in the j^{th} school eligible for free or reduced-price meals, centered at the grand mean; $SmallTown_j$ is a dummy variable taking the value 1 if the j^{th} school is in a small town and 0 otherwise, centered at the grand mean, $\overline{SmallTown}$; $LargeTown_j$ is a dummy variable taking the value 1 if the j^{th} school is in a large town or on the fringe of a city and 0 otherwise, centered at the grand mean, $\overline{LargeTown}$[72]; $Delta_j$ is a dummy variable taking the value 1 if the j^{th} school is in the Delta and 0 otherwise, centered at the grand mean, $\overline{Delta}$; and υ_{0j} is the error term for the j^{th} school (school-level error terms are assumed to be normally distributed with mean 0 and variance τ^2).

The parameter γ_{01} indicates K-PAVE's impact on a specified classroom outcome. Thus, a positive and significant estimate of γ_{01} indicates that there is compelling scientific evidence that the K-PAVE intervention influences classroom instructional practice. The statistical significance is assessed and the effect sizes of classroom impacts calculated following the same approach described for student impacts.

MODELS SPECIFICATION FOR GLOBAL *F*-TEST OF JOINT IMPACT ON MULTIPLE CLASSROOM INSTRUCTION OUTCOMES WITHIN A DOMAIN

The secondary confirmatory analysis examined three classroom instruction outcomes within a single domain: vocabulary and comprehension support, instructional support, and emotional support. All three involve instructional practices hypothesized to foster impacts on students' vocabulary and vocabulary-related outcomes. Before analyzing any of these measures individually, a joint test of the null hypothesis that the impact of K-PAVE on all three outcomes is 0 was conducted. The following ordinary least squares regression model was simultaneously estimated for all three outcomes within the domain. Values for all classroom-level variables (that is, posttest outcome, baseline outcome, and teacher covariates) were averaged for each school, and an ordinary least squares regression model

[72] As in the models testing impacts on students, the omitted category for locale is *RURAL*, which indicates that the school is in a rural area.

was estimated using school mean values and including weights to adjust for the loss of the school that dropped out of the study (as described above).

(K8)

$$\overline{Y_k} = \beta_0 + \beta_1 T_k + \beta_2 \overline{PRE_k} + \beta_3 \overline{tch\_female_k} + \beta_4 \overline{Tch\_AfrAm_k} + \beta_5 \overline{College_k} + \beta_6 \overline{GradDegree_k}$$
$$+ \beta_7 \overline{CertEC_k} + \beta_8 \overline{Cert\,\mathrm{Re}\,ad_k} + \beta_9 \overline{YrsTch_k} + \beta_{10} \overline{YrsTchKg_k}$$
$$+ \beta_{11} ReadingFirst_k + \beta_{12} MSStateInit_k + \beta_{13} AchLvlIndex_k + \beta_{14} PctAfrAm_k + \beta_{15} PctFreeLunch_k$$
$$+ \beta_{16} SmallTown_k + \beta_{17} L\arg eTown_k + \beta_{18} Delta_k + \varepsilon_k$$

where $\overline{Y_k}$ is an average value of a classroom outcome measure (for example, vocabulary and comprehension support) for the k^{th} school; T_k is a dummy variable for treatment status taking the value 1 for a school assigned to the K-PAVE treatment and 0 for a school assigned to the control group; $\overline{PRE_k}$ is the average baseline classroom instruction score for the k^{th} school; $\overline{tch\_female_k}$ is the proportion of teachers sampled from the k^{th} school that are female; $\overline{Tch\_AfrAm_k}$ is the proportion of teachers sampled from the k^{th} school that are African American; $\overline{College_k}$ is the proportion of teachers sampled from the k^{th} school whose highest level of education is a bachelor's degree; $\overline{GradDegree_k}$ is the proportion of teachers sampled from the k^{th} school whose highest level of education is a graduate degree; $\overline{CertEC_k}$ is the proportion of teachers sampled from the k^{th} school who have a teaching certificate in early childhood; $\overline{CertRead_k}$ is the proportion of teachers sampled from the k^{th} school with a teaching certificate in reading; $\overline{YrsTch_k}$ is the average number of years that teachers sampled from the k^{th} school have been teaching children (that is, total years of experience teaching); $\overline{YrsTchKg_k}$ is the average number of years that teachers sampled from the k^{th} school have been teaching kindergarten; $ReadingFirst_k$ is a dummy variable taking the 1 if the k^{th} school participated in the Reading First program in the 2008/09 school year and 0 otherwise; $MSStateInit_k$ is a dummy variable taking the value 1 if the k^{th} school has a Mississippi reading initiative and 0 otherwise; $AchLvlIndex_k$ is the Achievement Level Index for the k^{th} school; $PctAfrAm_k$ is the percentage of all students in the k^{th} school that are African American; $PctFreeLunch_k$ is the percentage of all students in the k^{th} school that are eligible for free or reduced-price meals; $SmallTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a small town and 0 otherwise; $L\arg eTown_k$ is a dummy variable taking the value 1 if the k^{th} school is in a large town or on the fringe of a city and 0 otherwise[73]; $Delta_k$ is a dummy variable taking the value 1 if the k^{th} school is in the Delta and 0 otherwise; β_0 is the covariate adjusted mean value of the outcome measure for control schools whose

[73] The reference category for locale is *RURAL*, which indicates that the school is located in a rural area.

covariate values are all 0; β_1 is the average treatment effect; β_2 through β_{18} are regression coefficients indicating the effect of each covariate on the outcome; and ε_k is the error term for the k^{th} school (the assumed distribution of these residuals is normal with mean 0 and variance $= \sigma^2$).

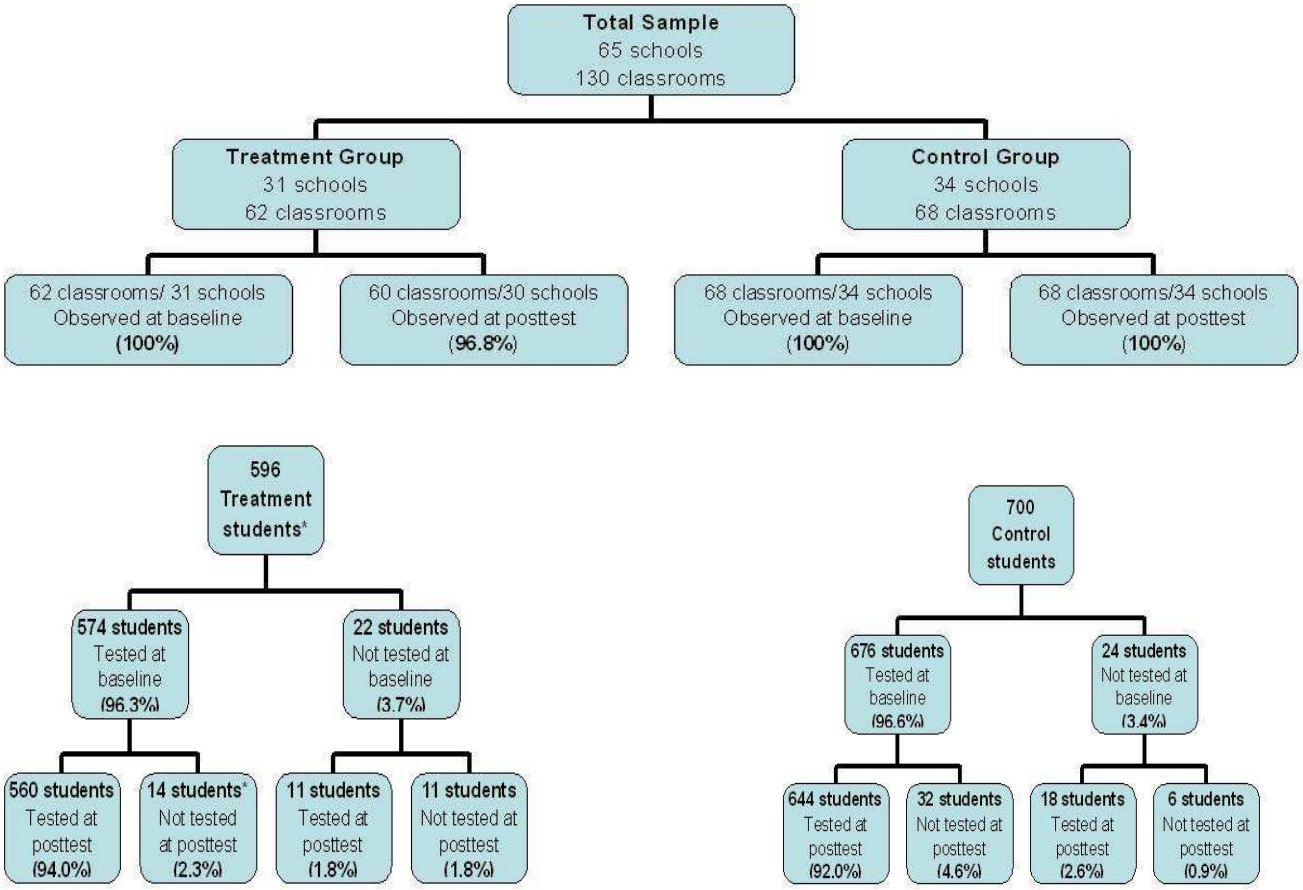
This model was estimated simultaneously for all three outcomes—vocabulary and comprehension support (*Voc_Comp*), instructional support (*Instr*), and emotional support (*Emot*)—in the domain, using three outcome equations. Correlations among error terms in the three equations were assumed to be 0. The equations were estimated simultaneously in order to be able to impose the null hypothesis that the treatment effect in all three equations is 0 and to be able to conduct a joint *F*-test of the hypothesis, which is stated as

(K9)

$$H_0 = \beta_{1(VOC\_COMP)} = 0 \text{ and } \quad H_0 = \beta_{1(INSTR)} = 0 \text{ and } H_0 = \beta_{1(EMOT)} = 0$$

If the *p*-value for this hypothesis test was less than .05, then the null hypothesis was rejected and a multilevel model (as described above) was estimated to test the impact of K-PAVE on each of the individual classroom instruction outcomes. If the *p*-value for this hypothesis test was greater than .05, then the null hypothesis was not rejected and the impact of K-PAVE on the individual classroom instruction outcomes is not estimated.

APPENDIX L. FLOWCHART ILLUSTRATING SAMPLE ATTRITION FROM DATA COLLECTION



Note: There were 23 students in the treatment school that dropped out of the study, and they are excluded from this figure and the analytic sample. With these 23 students, there were 619 students selected for the treatment group, and 594 were tested at baseline. In other words, 20 of the 23 students from the school that dropped were tested at baseline, and none was tested at posttest.

APPENDIX M. MISSING DATA IMPUTATION

Two approaches were employed to impute missing data: dummy variable adjustment for missing covariates other than pretest scores and single stochastic regression imputation for missing student assessments and classroom instruction measures. Missing data imputation was done separately for treatment and control groups. For each type of missing data (covariates, test scores), the extent of the missing data in the treatment group and in the control group is described first followed by the imputation approach used.

DUMMY VARIABLE ADJUSTMENT FOR MISSING COVARIATES

Missing data on student covariates

Data were collected for the following student covariates:

- Age at posttest (age)
- Gender
- Race
- Eligibility for free or reduced-price meals
- Special education status

In the treatment group 1.5% of students (9 of 596) were missing data on at least one of the covariates, and in the control group 0.7% of students (5 of 700) were.

Missing data on school covariates

Data were collected for the following school covariates:

- Reading initiative in place prior to K-PAVE (Reading First, a Mississippi state initiative, or other)
- Achievement Level Index[74]
- Percentage of students who are African American
- Percentage of students eligible for free or reduced-price meals
- Locale (rural, small town, or large town/fringe of city)
- Delta region (or neighboring county)

[74] The Achievement Level Index is created by the Mississippi Department of Education for each school in the state, based on student performance on the Mississippi state accountability test (that is, Mississippi Curriculum Test), which is administered to all students in grade 3 or higher. Scores at all grade levels and for all subject areas are included in the index. The percentage of students in the school who score basic or higher and the percentage of students in the school scoring proficient or higher are used to create an Achievement Level Index score ranging from 100 to 600, with scores in the 100 range corresponding to a school performance level of "low" and scores in the 500 range corresponding to a school performance level of "superior."

There were no missing data in either the treatment or control group on any of the school covariates except for Achievement Level Index (Table M1).

Table M1. Missing data on school covariates in treatment and control samples

	Treatment (30 schools)		Control (34 schools)	
Covariate	***n***	***%***	***n***	***%***
Reading initiative	0	0.0	0	0.0
Achievement Level Index	4	13.3	3	8.8
Percentage of students who are African American	0	0.0	0	0.0
Percentage of students eligible for free or reduced-price meals	0	0.0	0	0.0
Locale	0	0.0	0	0.0
Delta region	0	0.0	0	0.0

Missing data on teacher covariates

Data were collected on the following teacher covariates:

- Gender
- Race
- Highest level of education
- Number of years teaching
- Number of years teaching kindergarten
- Has teaching certification in early childhood education
- Has teaching certification in reading instruction

There were 3.1% of teachers (4 of 128 teachers) that were missing data on at least one covariate.

Applying the dummy variable adjustment

The dummy variable adjustment was applied to all missing student, school, and teacher covariates, except for missing student and classroom instruction pretest scores. In this approach, missing values were all set to a constant value of 0. In addition, the analysis included an indicator variable identifying observations for which the value of the covariate is missing. The indicator variable for a given covariate was set to one in cases where the covariate was missing and was set to 0 in cases where the covariate was not missing. While some articles have shown that this method will generally produce biased estimates (for example, Jones 1996), a recent National Center for Education Evaluation and Regional Assistance Technical Methods report found that this method performed well in simulations that mirrored the randomized study design and analysis plan of this evaluation (Puma, Olsen, Bell, & Price 2009).

This approach is described here using the missing data on gender (*GENDER*) as an example. A new variable, *GENDER_IMP*, was created, which was set to *GENDER* for all nonmissing cases and to 0 for all missing cases. In addition, a second new variable, *GENDER_IMP _FLAG*, was created, which was set to 1 for all students for whom *GENDER* was missing and to 0 for all students for whom *GENDER* was not missing. Both new variables were included in place of the original variable (*GENDER*) in the multilevel model used to estimate impacts of K-PAVE on students. The same approach was used for each student, teacher, and school covariate with missing values, and each pair of new variables was included in the impact model simultaneously in place of the original variables that had missing values. No distinctions were made between treatment and control group observations in applying this procedure.

SINGLE STOCHASTIC REGRESSION IMPUTATION FOR MISSING PRETEST AND POSTTEST DATA

Missing pretest and posttest data

The percentage of students missing pretest assessment data is 3.7%–6.0% in the treatment group and 3.4%–3.6% in the control group. The percentage of students missing posttest assessment data is 4.2%–4.4% in the treatment group and 5.4%–5.6% in the control group (Table M2).

Table M2. Missing data on student pretest and posttest assessments, for treatment and control groups (596 treatment students, 700 control students)

	Expressive Vocabulary Test–2nd Edition		Listening comprehension		Academic knowledge	
	Treatment group	Control group	Treatment group	Control group	Treatment group	Control group
Missing pretest only	11 (1.9%)	18 (2.6%)	25 (4.2%)	28 (4.0%)	11 (1.9%)	19 (2.7%)
Missing posttest only	14 (2.4%)	32 (4.6%)	14 (2.3%)	32 (4.6%)	15 (2.5%)	32 (4.6%)
Missing both	11 (1.9%)	6 (0.9%)	11 (1.8%)	7 (1.0%)	11 (1.9%)	6 (0.9%)
Total missing pretest	22 (3.7%)	24 (3.4%)	36 (6.0%)	25 (3.6%)	22 (3.7%)	25 (3.6%)
Total missing posttest	25 (4.2%)	38 (5.4%)	25 (4.2%)	39 (5.6%)	26 (4.4%)	38 (5.4%)

For the four measures of classroom instruction—vocabulary and comprehension support, instructional support, emotional support, and time on literacy areas other than vocabulary and comprehension—one posttest measure was missing for only one classroom. No other measures were missing.

Applying single stochastic regression imputation

For student assessments and the classroom instruction measure, missing pretest scores and missing posttest scores were both imputed using single stochastic regression[75] (see Puma, Olsen, Bell, & Price 2009 for a description). To impute missing values using single stochastic regression, a multiple regression model adjusted for the multilevel structure of the data was used to estimate predicted values for each pretest or posttest with missing values. Predictors in each imputation model included all other available information collected (including pretest scores, posttest scores, and covariates). For each pretest and posttest measure with missing values, an imputation model was estimated using cases with complete data. For each missing score, a randomly selected residual was added to the predicted value from the regression model to obtain an imputed value (that is, imputed value = predicted value + a randomly selected residual). The residual error was added to predicted values in an effort to have the same variation among imputed values as among observed values. As noted above, imputation models were estimated separately for treatment and control groups.

All known variables were included as predictors of variables with missing values. When imputing pretest scores, posttest scores were included on the same measure as predictors, as well as pretest and posttest scores for other student outcomes (or other classroom instruction outcome measures). For example, for students missing expressive vocabulary pretest scores, all covariates were used to obtain predicted values, and students' expressive vocabulary posttest scores were included as predictors of the missing pretest scores. Although it seems unusual to use a posttest score to impute a pretest score, which will in turn be used to predict the posttest outcome in the impact model, this approach is recommended by experts (Little & Rubin, 2002; Moons et al, 2006; and Allison, 2002, as cited in Puma et al, 2009).

In addition to using posttest scores on the same measure to predict a pretest with missing values, pretest and posttest scores for other student outcomes were also included. In the example above with students missing Expressive Vocabulary Test pretest scores, pretest and posttest measures of academic knowledge and listening comprehension were included as predictors in the imputation model, as were all the covariates and the Expressive Vocabulary Test posttest measures. Following the recommendations of Puma et al. (2009), the imputation models included any measured variables that may be associated with missing data.

Equation M1 shows the model used to predict missing student assessment pretests. In this example, the Expressive Vocabulary Test–2nd Edition pretest is predicted for students in treatment schools. The same approach was used for all student assessments. A series of 29 treatment school dummy variables were included in the model to adjust for the nesting of

[75] The literature suggests that in general single stochastic regression imputation produces standard error estimates that are biased downward and that this problem can be addressed by multiple stochastic regression imputation (see, for example, Allison 2002). However, Puma et al. (2009) found that when schools were randomized but data were missing at the student level, single stochastic regression imputation did not yield standard error estimates that were biased downward. Therefore, multiple imputation would seem to be unnecessary in this context.

students in 30 treatment schools.[76] The same model was estimated separately for control schools, with 33 control school dummy variables (for the 34 control schools) instead of the treatment school indicators.

(M1)

$$
\begin{aligned}
EVT\_pretest = {} & \beta_0 + \beta_1 EVT\_posttest \\
& + \beta_2 AcadKnow\_pretest + \beta_3 AcadKnow\_posttest \\
& + \beta_4 ListeningComp\_pretest + \beta_5 ListeningComp\_posttest \\
& + \beta_6 Stud\_PosttestAge + \beta_7 Stud\_Female \\
& + \beta_8 Stud\_AfrAm + \beta_9 Stud\_EligFreeReducedLunch + \beta_{10} Stud\_IEP \\
& + \beta_{11} Sch\_ReadingFirst + \beta_{12} Sch\_MSStateInit + \beta_{13} Sch\_AchLvlIndex \\
& + \beta_{14} Sch\_PctAfrAm + \beta_{15} Sch\_PctEligFreeReducedLunch \\
& + \beta_{16} SmTown + \beta_{17} LgTown + \beta_{18} Delta \\
& + \alpha_1 TreatmentSch1 + \ldots + \alpha_{29} TreatmentSch29 + \varepsilon
\end{aligned}
$$

This single stochastic regression imputation approach was also used to impute missing posttest data. As with missing pretest scores, imputation was done separately for treatment and control groups, and all covariates from the impact analysis model were used in the imputation model, as were pretest and posttest scores for other student assessments. Equation M2 shows the model used to predict missing student assessment posttest scores, using the Expressive Vocabulary Test posttest for the control group as an example. Again, 33 control school dummy variables were included in the model to adjust for the nesting of students in 34 control schools. The same model was used to predict missing posttest scores in the treatment group; however, 29 treatment school dummy variables were used in the model instead of the control school dummy variables.

(M2)

$$
\begin{aligned}
EVT\_posttest = {} & \beta_0 + \beta_1 EVT\_pretest \\
& + \beta_2 AcadKnow\_pretest + \beta_3 AcadKnow\_posttest \\
& + \beta_4 ListeningComp\_pretest + \beta_5 ListeningComp\_posttest \\
& + \beta_6 Stud\_PosttestAge + \beta_7 Stud\_Female \\
& + \beta_8 Stud\_AfrAm + \beta_9 Stud\_EligFreeReducedLunch + \beta_{10} Stud\_IEP \\
& + \beta_{11} Sch\_ReadingFirst + \beta_{12} Sch\_MSStateInit + \beta_{13} Sch\_AchLvlIndex \\
& + \beta_{14} Sch\_PctAfrAm + \beta_{15} Sch\_PctEligFreeReducedLunch \\
& + \beta_{16} SmTown + \beta_{17} LgTown + \beta_{18} Delta \\
& + \alpha_1 ControlSch1 + \ldots + \alpha_{33} ControlSch33 + \varepsilon
\end{aligned}
$$

The same approach was used to impute the one missing value on the vocabulary and comprehension support posttest measure (see equation M3). Predictors in the model were school covariates, teacher covariates, baseline vocabulary and comprehension support, other classroom instruction baseline and posttest measures, and a series of 33 control school

[76] A series of dummy variables was used to adjust for the nested structure of the data rather than estimating a multilevel model for imputing missing values.

dummy variables to adjust for the nesting of classrooms in 34 control schools. There was no imputation model for the treatment group, as there were no missing classroom data for the treatment group.

(M3)

$$
\begin{aligned}
VocCompSup\_posttest = {} & \beta_0 + \beta_1 VocCompSup\_pretest \\
& + \beta_2 InstrSup\_pretest + \beta_3 InstrSup\_posttest \\
& + \beta_4 EmotSup\_pretest + \beta_5 EmotSup\_posttest \\
& + \beta_6 OthLit\_pretest + \beta_7 OthLit\_posttest \\
& + \beta_8 Tch\_Female + \beta_9 Tch\_AfrAm + \beta_{10} College \\
& + \beta_{11} GradDegree + \beta_{12} CertEC + \beta_{13} Cert\,\mathrm{Re}ad \\
& + \beta_{14} YrsTch + \beta_{15} YrsTchKg + \beta_{16} Sch\_ReadingFirst \\
& + \beta_{17} Sch\_MSStateInit + \beta_{18} Sch\_AchLvlIndex \\
& + \beta_{19} Sch\_PctAfrAm + \beta_{20} Sch\_PctEligFreeReducedLunch \\
& + \beta_{21} SmTown + \beta_{22} LgTown + \beta_{23} Delta \\
& + \alpha_1 ControlSch1 + \ldots + \alpha_{33} ControlSch33 + \varepsilon
\end{aligned}
$$

Once all missing values were imputed for student assessment pretests and posttests and for the one missing classroom instruction posttest measure using the single stochastic regression approach, the student and classroom impact models were estimated using imputed values. As a form of sensitivity analysis, impact models were also estimated using only cases with nonmissing pretest and posttest data and with missing covariates imputed using the dummy variable method, using only cases with nonmissing covariates but with missing pretest and posttest data imputed using single stochastic regression, and using only cases with nonmissing data for pretests, posttests, and covariates (that is, listwise deletion). Sensitivity analyses (reported in Appendix N) indicate that the magnitude and standard errors of impact estimates were similar regardless of whether missing pretest data, missing posttest data, or missing covariate data were imputed. The imputation of missing data did not affect whether impact estimates were statistically significant.

APPENDIX N. SENSITIVITY ANALYSES

IMPACTS ON STUDENTS

Expressive vocabulary

Sensitivity analyses were conducted to examine the robustness of the impact estimates. The results of the sensitivity analyses were compared with the models presented in Chapter 4. For the Expressive Vocabulary Test–2nd Edition (EVT–2), the impact model reported in Chapter 4 is a three-level model (school, classroom, and student; see Appendix K on model specifications), with a treatment indictor included at the school level to estimate the average impact of K-PAVE on students' EVT–2 posttest score. The impact estimate is adjusted for students' baseline EVT–2 test score, other student covariates, and school characteristics. Missing values for pretest and posttest scores are imputed using single stochastic regression, and missing values for school and student covariates are imputed using the dummy variable adjustment.

Models were estimated to examine the sensitivity of the findings from the final impact model to covariate adjustment, missing data imputation, delayed baseline student testing, student crossovers, nonparticipation, students with a score of 0 on the EVT–2, and outliers. The sensitivity analysis models (described below) were compared with the final impact model in Chapter 4:

- One model testing sensitivity to covariate adjustment was estimated with no covariates other than the treatment indicator and EVT–2 pretest score. All other models estimated as part of the sensitivity analyses had the same structure and covariates as the final model.
- Four models testing sensitivity to missing data imputation were estimated:
    - without imputing any missing values—posttest, pretest, or covariates (that is, with a sample including only complete cases, or listwise deletion), without imputing missing values for the EVT–2 posttest and pretest (that is, with a sample including only students with both posttest and pretest scores) but using the dummy variable method for missing covariates without imputing missing values for covariates other than pretest scores (that is, with a sample including only students with no missing covariates) but imputing missing posttest values using single stochastic regression imputation, and without imputing scores for incorrectly administered tests and instead dropping these cases from the analysis.
    - Two models testing sensitivity to delayed baseline testing were estimated. One excluded 29 students tested more than one week after K-PAVE training was completed, and the other excluded only the 12 students tested three weeks after all other student testing was completed.
- One model testing sensitivity to crossovers was estimated and excluded the five students that transferred to another study school and crossed conditions, for whom posttest data were available.
    - Two models testing sensitivity to movement to non-study schools were estimated. One excluded 17 students—11 from the treatment group and 6 from the control group—that transferred to another school and were not tested at pretest and not tested at posttest, and the other excluded 45 students—21 from the treatment group and 24 from the control

group—that transferred to another school and were not tested at posttest (this group includes the 17 students that were also not tested at pretest).

- One model testing sensitivity to scoring 0 on the EVT pretest was estimated and excluded students who had a 0 score.
- One model testing sensitivity to outliers was estimated and excluded six outliers identified in Appendix V. For the EVT–2, observations with studentized residuals that were more than four standard deviations from the mean were considered outliers.
- One model testing sensitivity to weighting schools to adjust for the school that dropped out was estimated.

In all models, impact estimates from the sensitivity analysis models are statistically significant, with magnitudes ranging from 1.2 to 1.7, standard errors ranging from 0.56 to 0.64, and standardized effect sizes ranging from 0.11 to 0.15 (Table N1). Findings were found to be robust.

Other vocabulary-related student outcomes: academic knowledge and listening comprehension

To minimize the amount of type I error introduced by multiple hypothesis testing, a global *F*-test was conducted of the null hypothesis that the impact of K-PAVE on the two vocabulary-related student tests – academic knowledge and listening comprehension – was 0 for both outcomes. Impacts on each of the individual outcomes measured within a domain were examined only if we were able to reject the null hypothesis of no impact on both outcomes, which was tested by the global test.

The approach to conducting the global *F*-test involved simultaneously estimating ordinary least squares regression models—one for academic knowledge and one for listening comprehension—in which school mean posttest scores were regressed on school mean pretest scores, school means for student covariates, and school covariates (see Appendix K on model specifications). In addition, school means were weighted to reflect the number of students sampled in each school.

Table N1. Estimated impact on kindergarten students' expressive vocabulary (EVT–2) in models fit for sensitivity analysis compared with final impact model in Chapter 4.

	Impact estimate	Standard error	*t*-statistic	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	1.60	0.585	2.74	.006	0.43	2.77	0.141
Covariate adjustment							
No covariates except treatment status and pretest	1.19	0.595	1.99	.05	–0.004	2.38	0.105
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	1.45	0.624	2.32	.02	0.20	2.70	0.128
Excluding cases with missing test scores	1.67	0.568	2.94	.003	0.53	2.81	0.147
Excluding cases with missing covariates (other than pretest)	1.47	0.643	2.29	.02	0.19	2.76	0.130
Excluding cases with incorrectly administered tests	1.55	0.588	2.64	.008	0.38	2.73	0.137
Delayed baseline testing							
Excluding 27 students with baseline testing at least one week late	1.74	0.576	3.02	.003	0.58	2.87	0.153
Excluding 12 students with baseline testing three weeks late	1.63	0.580	2.80	.005	0.47	2.79	0.143
Crossovers and out-migrants							
Excluding five crossovers	1.66	0.592	2.81	.005	0.48	2.85	0.147
Excluding transfers missing pre- and posttest	1.66	0.580	2.85	.004	0.49	2.82	0.146
Excluding transfers missing posttest	1.66	0.580	2.86	.004	0.50	2.82	0.146
Zero raw score							
Excluding students with zero raw score	1.52	0.568	2.68	.007	0.39	2.66	0.134
Outliers							
Excluding six outliers	1.43	0.556	2.57	.01	0.32	2.54	0.126
Weighting schools							
Excluding school weights adjusting for school drop out	1.60	0.585	2.74	.006	0.43	2.77	0.141

Sensitivity analyses were conducted to examine whether the findings from the global *F*-test were robust. Models were estimated to examine the sensitivity of the findings to the same factors examined for the EVT: covariate adjustment, missing data imputation,[77] delayed baseline student testing, student crossovers, out-migration, and students with a score of 0 on the academic knowledge test.[78] The sensitivity of findings to outliers was not tested because no outliers were identified. Results of the sensitivity analyses are presented in Table N2. Findings were found to be sensitive to the procedures used to address imputation of incorrectly administered baseline tests, delayed baseline testing, student crossovers, imputation of values for students transferring out of study schools, and students with a 0 posttest raw score. In 7 of 10 models estimated by varying these factors, the null hypothesis of 0 impact on both listening comprehension and academic knowledge was rejected (see Table N2). For this reason, the impact of K-PAVE on listening comprehension and academic knowledge was estimated separately using a Bonferroni correction to limit the heightened type I error associated with conducting multiple comparisons within a domain.

Table N2. Results of sensitivity analysis conducted on joint test of K-PAVE impact on academic knowledge and listening comprehension

	***F*-statistic**	***p*-value**
Final model (reported in Chapter 4)	$F_{2,96} = 3.08$	.0508
Covariate adjustment		
No covariates except treatment status and pretest	$F_{2,122} = 2.48$	.09
Missing data imputation		
Posttest and pretest school means calculated without missing values	$F_{2,96} = 2.87$	.06
School covariates not imputed and posttest and pretest school means calculated without missing values (7 of 64 schools lost)	$F_{2,84} = 1.69$	.19
Excluding cases with incorrectly administered tests from school means	$F_{2,96} = 3.24$	.04
Delayed baseline testing		
Excluding 27 students with baseline testing at least one week late	$F_{2,94} = 4.57$	.01
Excluding 12 students with baseline testing three weeks late	$F_{2,96} = 3.53$	.03
Crossovers and out-migrants		
Excluding five crossover students from school means	$F_{2,96} = 3.25$	.04
Excluding student transfers from school means: missed pretest and posttest	$F_{2,96} = 3.56$	.03
Excluding student transfers from school means: missed posttest	$F_{2,96} = 3.77$	.03
Zero raw score		
Excluded from school mean, students with a 0 academic knowledge raw score	$F_{2,96} = 3.21$	.045

Results of these sensitivity analyses are presented in Table N3 for the model testing the impact of K-PAVE on listening comprehension and in Table N4 for the model testing the

[77] In the models used to conduct the global *F*-test, missing values for student covariates were not imputed using the dummy variable adjustment but were based only on nonmissing values for the school. School means for student pretest and posttest scores were calculated with missing values imputed using single stochastic regression, and sensitivity analyses included school means calculated based only on nonmissing pretest and posttest values. For missing school covariates, the dummy variable adjustment was applied, and sensitivity analysis examined models excluding schools with missing values for school covariates.

[78] Raw scores of 0 on the listening comprehension test were common; thus we did not examine sensitivity of findings to the presence of 0 scores on the listening comprehension test.

impact on academic knowledge. Findings regarding listening comprehension were found to be robust (see Table N3). Impacts were not found to be statistically significant in any models. Unstandardized impact estimates ranged from 1.0 to 1.6, with standard errors ranging from 0.84 to 0.94 and standardized effect sizes ranging from 0.08 to 0.12.

The positive and statistically significant impact of K-PAVE on academic knowledge was also found to be robust in the sensitivity analysis (see Table N4). Across all sensitivity models, the impact estimates ranged from 1.8 to 2.2, standard errors ranged from 0.81 to 0.90, and standardized effect sizes ranged from 0.13 to 0.16. However, using the 0.025 threshold for statistical significance set by the Bonferroni adjustment, the impact was not statistically significant in 4 of the 12 sensitivity models. The *p*-values range from.026 to .043, which indicates some sensitivity to covariate adjustment, imputation of missing covariate data (two models), and imputation of scores for tests that were incorrectly administered. Nonetheless, the magnitude of impact estimates and standard errors are similar to those in the main model. In three of the four models, the impact estimate ranges from 1.84 to 2.02, compared with 1.95 in the main model, with standard errors ranging from 0.89 to 0.90, compared with 0.85. Effect sizes for all four sensitivity models with *p*-values between .025 and .05 were .13 to .14. The impact estimate was lower in the model without covariate adjustment, at 1.71. Although these four models had *p*-values above .025, the preponderance of the evidence indicates an impact occurred, even when judged by the conservative standards of the Bonferroni test for significance when dealing with multiple comparisons.

Table N3. Estimated impact on kindergarten students' listening comprehension in models fit for sensitivity analysis compared with final impact model presented in Chapter 4

	Impact estimate	Standard error	*t*-statistic	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	1.41	0.881	1.60	.11	–0.35	3.17	.109
Covariate adjustment							
No covariates except treatment status and pretest	1.00	0.845	1.19	.23	–0.69	2.69	0.077
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	1.39	0.937	1.48	.14	–0.48	3.27	0.107
Excluding cases with missing test scores	1.46	0.865	1.69	.09	–0.28	3.19	0.112
Excluding cases with missing covariates (other than pretest)	1.32	0.939	1.41	.16	–0.55	3.20	0.102
Delayed baseline testing							
Excluding 27 students with baseline testing at least one week late	1.57	0.877	1.79	.07	–0.18	3.33	0.121
Excluding 12 students with baseline testing three weeks late	1.51	0.869	1.74	.08	–0.22	3.25	0.117
Crossovers and out-migrants							
Excluding five crossovers	1.38	0.882	1.56	.12	–0.39	3.14	0.106
Excluding transfers missing pretest and posttest	1.44	0.872	1.65	.10	–0.30	3.18	0.111
Excluding transfers missing posttest	1.47	0.844	1.74	.08	–0.22	3.15	0.113
Weighting schools							
Excluding school weights adjusting for school drop out	1.42	0.90	1.61	.11	-0.38	3.22	0.109

Table N4. Estimated impact on kindergarten students' academic knowledge in models fit for sensitivity analysis compared with final impact model presented in Chapter 4

	Impact estimate	Standard error	*t*-statistic	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	1.95	0.851	2.29	.022	0.25	3.65	.144
Covariate adjustment							
No covariates except treatment status and pretest	1.71	0.819	2.08	.037	0.07	3.35	0.127
Missing Data Imputation							
Excluding cases with any missing data (listwise deletion)	1.99	0.892	2.23	.026	0.21	3.78	0.148
Excluding cases with missing test scores	2.01	0.824	2.44	.015	0.36	3.65	0.149
Excluding cases with missing covariates (other than pretest)	2.02	0.903	2.24	.026	0.21	3.83	0.150
Excluding cases with incorrectly administered tests	1.84	0.904	2.03	.043	0.08	3.65	0.136
Delayed baseline testing							
Excluding 27 students with baseline testing at least one week late	2.19	0.813	2.70	.007	0.57	3.82	0.163
Excluding 12 students with baseline testing three weeks late	2.05	0.820	2.50	.013	0.41	3.69	0.152
Crossovers and out-migrants							
Excluding five crossovers	1.98	0.853	2.32	.020	0.28	3.69	0.147
Excluding transfers missing pretest and posttest	1.99	0.847	2.35	.019	0.29	3.68	0.147
Excluding transfers missing posttest	1.93	0.835	2.31	.021	0.26	3.60	0.143
Zero raw score							
Excluding students with 0 raw score	2.02	0.868	2.32	.020	0.28	3.75	0.150
Weighting schools							
Excluding school weights adjusting for school drop out	1.95	0.85	2.29	.022	0.24	3.65	0.144

IMPACTS ON CLASSROOM INSTRUCTION

Sensitivity analyses were conducted to examine the robustness of the estimates of impacts on classroom instruction outcomes. The domain of classroom instructional practices hypothesized to foster impacts on students' vocabulary was measured by three outcomes: vocabulary and comprehension support, instructional support, and emotional support. Sensitivity analyses were conducted on the models estimated to conduct the global *F*-test, which tested the null hypothesis that impacts on all three types of instructional support were 0. Three ordinary least squares regression models—one for each type of instructional support—were estimated, in which school mean posttest scores were regressed on school mean pretest scores,[79] school means for teacher covariates,[80] and school covariates[81] (see Appendix K on model specifications). Because data were from the same number of classrooms (two) in each school, school means did not need to be weighted. The sensitivity of the findings to covariate adjustment and missing data imputation were examined by fitting four models and comparing the results of their global *F*-tests to the test from the model reported in Chapter 4:

- Models estimated with no covariates other than the treatment indicator and baseline measure of the outcome.
- Models estimated with all covariates but without imputation of missing values for school covariates (that is, with a sample including only the 57 schools with no missing school covariate data).
- Models estimated with all covariates, but school means for posttest[82] vocabulary and comprehension support were calculated without imputation of the one missing value.[83]
- Models estimated with all covariates, but without imputation of missing school covariates and with school posttest means calculated without imputation of one missing value.

The findings from the sensitivity analyses for the joint test of impacts on classroom instructional practices were found to be robust (Table N5). For all models the null hypothesis that the impacts on all three instructional practices hypothesized to foster students' vocabulary were 0 was rejected.

[79] There were no missing values for pretest scores, so school means were calculated based on complete cases.
[80] School means for teacher covariates were calculated based on non-missing values only.
[81] Missing values for school covariates were imputed using the dummy variable adjustment.
[82] No other classroom instruction pretest and posttest measures had missing values.
[83] The sample size of 64 schools is not affected by the inclusion or exclusion of the single imputed posttest value for the vocabulary and comprehension support composite. When the imputed value is included, the school mean is calculated based on posttest scores for 64 schools. When the imputed value is excluded, the school mean is calculated based on posttest scores for 63 schools.

Table N5. Results of sensitivity analysis conducted on joint test of K-PAVE impact on vocabulary and comprehension support, instructional support, and emotional support

	F-statistic	*p*-value
Final model (reported in Chapter 4)	$F_{3,132} = 5.26$	.002
Covariate adjustment		
No covariates except treatment and pretest	$F_{3,183} = 8.97$	< .0001
Missing data imputation		
School covariates not imputed	$F_{3,114} = 5.14$	.002
Vocabulary and comprehension support posttest mean calculated without imputing missing value	$F_{3,132} = 5.19$	.002
School covariates not imputed and vocabulary and comprehension posttest mean calculated without missing value	$F_{3,114} = 5.09$	.002

Because the null hypothesis was rejected, impacts on each of the three instructional practices—vocabulary and comprehension support, instructional support, and emotional support—were examined. As reported in Chapter 4, for each outcome and for the fourth classroom instruction outcome—proportion of observation cycles spent on non-vocabulary literacy instruction—a two-level model (classrooms, schools) was estimated, with a treatment indicator included at the school level to estimate the average impact on the classroom instruction outcome measure (see Appendix K on model specifications). The impact estimate was adjusted for baseline classroom instruction score, teacher characteristics, and school characteristics. Missing values for covariates other than the pretest were imputed using the dummy variable adjustment, and the one missing posttest value (vocabulary and comprehension support score for one classroom) was imputed using single stochastic regression.

The results of the sensitivity analyses were compared with those from the final models reported in Chapter 4 that examine impacts on classroom instruction. Models were estimated to examine the sensitivity of the findings from the impact models to covariate adjustment, missing data imputation, and weighting to adjust for school drop out. The same sensitivity analyses were conducted for all four classroom instruction measures—the three types of instructional practices hypothesized to foster students' vocabulary (vocabulary and comprehension support, instructional support, and emotional support) and the amount of time spent on non-vocabulary literacy instruction. For all four models, the results from three sensitivity analysis models were compared with those from the final models: one model estimated with no covariates other than the treatment indicator and baseline measure of the outcome; another model estimated with all covariates but without imputation of any missing covariates;[84] and a third model that was the same as the final model but without weights to adjust for school drop out.

[84] One school covariate was missing for 7 schools, reducing the sample size to 57 schools for models without imputation of missing school covariates. For missing teacher covariates and for the one classroom missing vocabulary and comprehension support posttest, excluding imputed values from the calculation of school means did not alter the sample size of schools.

In addition, for vocabulary and comprehension support, which was missing one posttest value, two additional models examining the sensitivity of the impact finding to missing data imputation were estimated:

- A model with all covariates but without imputation of the missing posttest score.
- A model with all covariates but without imputation of either missing covariates or the missing posttest score.

For all models of the sensitivity analyses for vocabulary and comprehension support, the null hypothesis that the impact on vocabulary and comprehension support was 0 was rejected (Table N6). The results were found to be robust. In the model with no teacher and school covariates, the impact estimate was higher, with an unstandardized impact of 0.85 compared with 0.73 (on a scale with a mean of 0 and a standard deviation of 1) and a standardized effect size of 0.97 compared with 0.82.

The results from the sensitivity analyses for instructional support were found to be sensitive to covariate adjustment and the imputation of missing values for covariates (Table N7). In both models the magnitude of the impact estimate was higher, with unstandardized impacts of 0.76 and 0.71, respectively, compared with 0.60 (on a 1–7 scale) and standardized effect sizes of 0.59 and 0.55, respectively, compared with 0.47. The standard errors were smaller[85] (0.31 and 0.33, respectively, compared with 0.36), and the impacts were found to be statistically significant.

The results from the sensitivity analysis for emotional support were found to be robust, and the null hypothesis of no impact on emotional support in any of the models could not be rejected (Table N8). The unstandardized impact ranged from 0.16 to 0.22, and the standardized effect size ranged from 0.20 to 0.27.

The results from the sensitivity analysis for time spent on nonvocabulary literacy instruction were found to be robust, and the null hypothesis of no impact on the proportion of observation cycles spent on nonvocabulary literacy instruction could be rejected in any of the models (Table N9). The unstandardized impact ranged from –0.01 to 0.02, and the standardized effect size ranged from –0.03 to 0.07.

[85] Although we would expect the covariates to increase the precision of the impact estimate, models examining K-PAVE impacts on classroom instruction that include covariates consistently have higher standard errors for the impact estimate than models without covariates (other than the baseline measure of the outcome). The standard errors increase with the inclusion of covariates because of covariates that—in the data (though not in expectation)—are correlated with treatment status. Random assignment of schools usually results in a sample that is balanced on school characteristics. However, as noted in chapter 2, a greater percentage of control schools than treatment schools are located in rural areas. The inclusion of covariates that are correlated with treatment status, although required to adjust for the imbalance between groups, causes the standard error of the impact estimate to increase.

Table N6. Results of sensitivity analyses of estimated K-PAVE impact on vocabulary and comprehension support

	Estimate	Standard error	*t*-stat	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	0.726	0.208	3.50	.0009	0.31	1.14	0.823
Covariate adjustment							
No covariates except treatment status and pretest	0.853	0.191	4.46	<.0001	0.47	1.24	0.967
Missing data imputation							
Missing covariates and posttest not imputed	0.699	0.190	3.67	.0006	0.32	1.08	0.793
Missing covariates not imputed	0.701	0.190	3.68	.0006	0.32	1.08	0.794
Missing posttest not imputed	0.726	0.207	3.51	.0009	0.31	1.14	0.822
Weighting schools							
Excluding weights adjusting for school drop out	0.727	0.201	3.50	.0009	0.32	1.13	0.824

Table N7. Results of sensitivity analyses of estimated K-PAVE impact on instructional support

	Estimate	Standard error	*t*-stat	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	0.60	0.36	1.69	.097	–0.11	1.32	0.470
No covariates except treatment status & pretest	0.758	0.31	2.46	.0169	0.14	1.36	0.592
Final model without imputing missing covariates	0.705	0.33	2.14	.04	0.05	1.36	0.551
Excluding weights adjusting for school drop out	0.60	0.36	1.68	.10	-.11	1.32	0.470

Table N8. Results of sensitivity analyses of estimated K-PAVE impact on emotional support

	Estimate	Standard error	*t*-stat	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	0.156	0.200	0.78	.44	–0.24	0.56	0.196
No covariates except treatment status and pretest	0.194	0.178	1.09	.28	–0.16	0.55	0.243
Final model without imputing missing covariates	0.184	0.182	1.01	.32	–0.18	0.55	0.230
Excluding one outlier	0.217	0.193	1.12	.27	–0.17	0.60	0.272
Excluding weights adjusting for school drop out	0.157	0.200	0.78	.44	-0.24	0.56	0.196

Table N9. Results of sensitivity analyses of estimated K-PAVE impact on proportion of cycles spent on nonvocabulary literacy instruction

	Estimate	Standard error	*t*-stat	*p*-value	95% confidence interval		Effect size
Final model (reported in Chapter 4)	–0.0107	0.074	–0.14	.89	–0.16	0.137	–0.03
No covariates except treatment status and pretest	0.0159	0.069	0.23	.82	–0.12	0.154	0.048
Final model without imputing missing covariates	–0.0025	0.066	–0.04	.97	–0.14	0.130	–0.008
Excluding weights adjusting for school drop out	-0.0108	0.074	–0.15	.88	-0.16	0.137	–0.032

APPENDIX O. SCHOOL, TEACHER, AND STUDENT COVARIATES

Student-level covariates used in the analysis are defined in Table O1.[86]

Table O1. Student-level covariates

Covariate	How measured
Score on student outcome measure at baseline (*Pre*)	Baseline **standardized score** on standardized student outcome measure • Expressive Vocabulary Test–2 • Academic knowledge • Listening comprehension
Female (*female*)	1 =student is female 0 =student is male
Student has an Individualized Education Plan (*StudentIEP*)	1 = student has an Individualized Education Plan and receives special education services for speech/language, occupational therapy, physical therapy, social skills, reading/literacy, mathematics, behavior modification, adaptive behaviors, self-care skills or other special educational needs 0 = student does not have an Individualized Education Plan
Eligibility for school meal program (*FreeReducedLunch*)	1 = student is eligible for free or reduced-price meals 0 = student is not eligible for free or reduced-price meals
African American (*AfricanAmerican*)	1 = student is African American 0 = student is not African American

[86] We planned to include two other covariates in the analysis: whether students had been retained in kindergarten and whether students had attended preschool prior to entering kindergarten. However, issues of data quality precluded their use. In the case of data on kindergarten retention, schools reported inconsistent information. Some schools reported whether students had been retained in the previous year (that is, whether they were attending their second year of kindergarten); other schools reported whether students would be retained in the current school year (that is, whether they would be attending a second year of kindergarten in the subsequent year). Because we did not have comparable data on all students, kindergarten retention status could not be included in the analysis. Schools also reported inconsistent information regarding whether students attended preschool prior to kindergarten. Some schools reported information about the type of arrangement that children attended prior to kindergarten (such as prekindergarten, Head Start, daycare, family childcare, public/private), but others reported either "yes" or "no," making it unclear what types of arrangements were included in the "yes" category and what types were not. For example, some schools may have included family childcare in their definition of attending preschool, while others may have included only prekindergarten and Head Start, excluding even programs in daycare centers. Furthermore, for 9.3% of the sample, schools reported that they did not know whether students attended preschool or they did not collect information on preschool. For another 10.2% of students, schools did not report preschool information and did not indicate why. Given that the information reported was not consistently defined and that no information was reported for 19.5% of students, data on preschool attendance were not included as a covariate.

Teacher characteristics—defined in Table O2—were averaged for each school and treated as covariates in the school-level model.

Table O2. Teacher-level covariates

Covariate measure	How measured
Female (*Tch_Female*)	1 = teacher is female 0 = teacher is male
African American (*Tch_AfrAm*)	1 = teacher is African American 0 = teacher is not African American (for this sample, teacher is White)
Level of education	Teacher's highest level of education completed, which is represented as a series of dummy variables, where 1 = specified level is highest level of education and 0 = specified level is not highest level of education. The reference category is "some graduate courses" (*SomeGrad*). • Bachelor's degree (*College*) • Some graduate courses (*SomeGrad*) • Graduate degree (*GradDegree*)
Teaching certification in early childhood (*CertEC*)	1 = teacher has a teaching certificate in early childhood education 0 = teacher does not have a teaching certificate in early childhood education
Teaching certification in reading instruction (*CertRead*)	1 = teacher has a teaching certificate in reading instruction 0 = teacher does not have a teaching certificate in reading instruction
Years teaching (*YrsTch*)	Number of years that the teacher has been teaching children. Values are on a continuous scale.
Years teaching kindergarten (*YrsTchKg*)	Number of years that the teacher has been teaching kindergarten. Values are on a continuous scale.

School-level covariates used in the analysis are defined in Table O3.

Table O3. School covariates

Covariate measure	How measured
Treatment status (*T*)	1 = school was randomly assigned to receive the K-PAVE treatment 0 = school was randomly assigned to the control group
Reading initiatives	Represented in the model by a series of two dummy variables. The reference category is *OthRead*. • Reading First (ReadingFirst): o 1 = school participates in the Reading First program o 0 = school does not have Reading First program • State reading initiative (MSStateInit) o 1 = school has a state reading initiative (for example, Barksdale, Reading Sufficiency) o 0 = school does not have a state reading initiative • Other (*OthRead*) o 1 = school has a reading program/curricula other than Reading First or a state reading initiative o 0 = school has either Reading First or a state reading initiative
Achievement Level Index (*AchLvlIndex*)	Measure of school-level achievement based on average scores on the Mississippi Curriculum Test, given annually to all students in grades 3 and higher. Scores are on a continuous scale from 100 to 600.
Percentage of students who are African American (*PctAfrAm*)	Percentage of students at school who are African American. Values range from 6% to 100%.
Percentage of students eligible for school meal program (*PctFreeLunch*)	Percentage of students at the school who are eligible for free or reduced-price meals. Values range from 40% to 100%.
Locale	Represented by a series of two dummy variables. The reference category is *Rural*. • Rural (*Rural*) o 1 = school is in a rural area o 0 = school is not in a rural area • Small town (*SmallTown*) o 1 = school is in a small town o 0 = school is not in a small town • Large town/fringe of city (*LargeTown*) o 1 = school is in a large town or on the fringe of a city o 0 = school is not in a large town or on the fringe of a city
Delta region (*Delta*)	1 = school is in the Delta region 0 = school is in a county contiguous to the Delta region

APPENDIX P. LIST OF K-PAVE MATERIALS PROVIDED TO TEACHERS

- 48 children's books
- Teacher guide
    - Description of three K-PAVE components (Building Bridges, Interactive Book Reading ["CAR Talk"], Explicit Vocabulary Instruction) and individual instructional strategies.
    - Template for weekly K-PAVE lesson plans.
    - Suggested tracking tools for monitoring conversations and reading with individual children.
    - Suggested schedule for integrating small group activities.
    - Instructions for creating vocabulary units on own.
- 24 K-PAVE teaching units
    - List of 10 target words for each unit to be posted in the classroom.
    - Quick definitions for 240 target words.
    - Sample CAR Talk questions for each book in the unit.
    - Brief description of two suggested extension activities per unit.
    - 360 laminated picture cards with pictures of 240 target words and 120 common words for use in Novel-Name Nameless-Category activity.

APPENDIX Q. SAMPLE WEEKLY UNIT FROM K-PAVE PROGRAM

TRANSPORTATION

Target Vocabulary

cargo	**helicopter**	**motor**	**pedal**
submarine	**oar**	**taxi**	**sailboat**
tire	**scooter**		

Books

Morris, A. (1990). *On the go.* New York: Scholastic Inc.

Ziefert, H. (2005). *From Kalamazoo to Timbuktu.* Maplewood, NJ: Blue Apple Books.

Quick definitions

On the Go

cargo things that are carried by a ship or an airplane

helicopter it flying machine with long blades that spin around on top

motor a machine that makes things go

pedal a part of a bicycle, car, or piano that you push with your foot

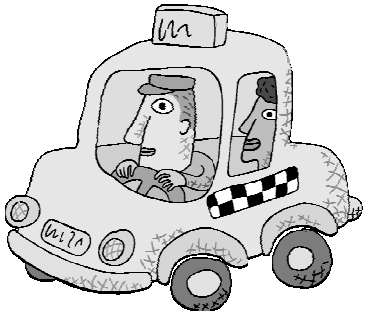
From Kalamazoo to Timbuktu

submarine a boat that goes under the water

oar a flat piece of wood you use to steer boats

sailboat a boat that that uses the wind to moves with

tire wheels on a car, truck or bicycle



Additional words

scooter something that has two wheels and a tall handle

you ride on **taxi** a car that you pay someone to take you somewhere

CAR TALK

On the Go

"Wheels make things go easier and faster. They can be *pedaled* or pushed. . ."

Competence: Where are the *pedals* on this bike?

Relate: When have you used *pedals*? What sorts of things have you *pedaled*?

"Or people. Some wheels are powered by *motors*?"

Abstract: What do you think the *motor* does?

"All aboard! Trains switch from *track* to *track*."

Competence: Where are the *tracks* for the train? Where are the tracks for the trolley? What about the monorail?

Abstract: How are all these *tracks* the same? How are they different?

"You can go straight up in a *helicopter* or a rocket. . . Liftoff!"

Relate: Has anyone seen a *helicopter* before? What does it look like? Would you ever want to take a ride in a *helicopter*?

From Kalamazoo to Timbuktu

"The bus blew *tires* in Butte, Montana. Millie took out her red bandana."

Competence: What happened to the bus *tires*?

Competence: What do you think the bus *tires* are filled with?

Abstract: What might make the *tires* of a bus to blow out?

Relate: Have you ever been on a bus or in a car when a *tire* popped? What happened?

"They set out across the Pacific Ocean with cases of food and suntan lotion."

Relate: Would you want to use an *oar* to get all the way across the ocean?

Abstract: Do you think it would be easier for them to use a *sailboat*?

PAIRED WORDS FOR TRANSPORTATION

Find the pictures for all the target words listed for this unit as well as the two words that are paired for each target word. The "paired words" are words that are known by the children. In the large group setting show all 3 pictures to the children and ask them which one is the "Target Word." This helps "map" the new or "novel" word to the unknown object. That is, children can begin to associate the new word with a new object or picture of an object.

Target words	Paired N^{3}C words
Cargo	Book, Circle
Helicopter	Bike, Truck
Motor	Train, Swing
Oar	Carrot, Saw
Pedal	Sock, Hand
Sailboat	Swing, Plane
Scooter	Orange, Jump
Submarine	Kitchen sink, Hammer
Taxi	Bed, Train
Tire	Apple, TV

EXTENSION ACTIVITIES

Activity 1: Sorting transportation by characteristics

Materials:

Pictures cards of types of transportation (*e.g., helicopter, scooter, submarine, taxi*)

Pictures cards of words associated with different types of transportation (*e.g., carries cargo, has pedals, has motors, has oars)*

•

Description:

The purpose of the activity is to have children practice discussing the attributes of various modes of transportation and sorting by the presence or absence of the attribute. Have children sort the transportation pictures into each classification (i.e., does it carry *cargo*? does it have *pedals*, does it have wheels, etc). Have children carry it out in pairs or as a group. They should talk together about what they are doing and why they are doing it. Repeat a number of times using each of the sorting classifications. If there is time, children can to offer their own ideas of attributes by which they could sort the transportation by (e.g., has seats, windows, goes fast, slow, etc.)

Activity 2: How would you get there from here?

Materials:

Transportation pictures from Activity 1.

Description:

Talk with children about the kinds of transportation they have used to get from one place to another and about the kinds of transportation they have seen in their communities. Show students the picture cards for transportation. Have children in the group take turns thinking of a destination to which he or she would like to go. Explain to children that we will be looking at the different ways that we get from one place to another. Make sure they understand that different children in the group might use different modes of transportation to get from one place to another. Place the pictures of modes of transportation within reach so that children can use them if they need to refer to them. Have children discuss the travel destination offered by the child and whether they would like to go there, too. Have them discuss how they might get from here to the destination. Some destinations might require multiple modes of transportation, so encourage the children to consider all the different types of transportation they might use.

APPENDIX R. LIST OF THE 240 K-PAVE TARGET WORDS

angry	castle	fireplace	letter	plantation	star
antennae	cave	fishing rod	library	plastic	steam
ants	cavity	flames	lighthouse	plow	stem
aphid	cello	flood	lightning	poison	submarine
appear	centers	floss	liquid	pole	supplies
arch	chick	flour	litter	police officer	sword
architect	chipmunk	flute	locomotive	praying mantis	tack
armor	cinnamon	foal	lunchbox	proud	tambourine
artist	clarinet	forecast	lure	puddle	tarantula
athlete	claws	French horn	magnet	radish	taxi
atlas	cliff	frustrated	map	rake	temperature
attract	cloud	fur	mask	recipe	thermometer
backpack	compass	garden	measuring cup	recycling bin	thermos
bacteria	compost	globe	melt	repel	tide pool
bait	conductor	gold	mitten	rise	tire
baker	confused	gravel	mole	root	track
barge	container	gum	monsoon	route	trench
barn	cork	gymnasium	moon	sailboat	trestle
battery	crossing guard	hall	motor	saliva	trombone
bay	cub	harp	mouthwash	sand dune	trumpet
beetle	cucumber	hatch	mower	scooter	tunic
blackberry	delta	hay	needle	seashore	tunnel
boil	desert	helicopter	nest	seeds	twig
bored	desk	helmet	nozzle	shade	veterinarian
boulder	disappear	highway	nurse	shed	volcano

braces	disappointed	hoe	oar	shell	waist
branch	disgusted	hook	ocean	shovel	wasp
bright	dragon	horseshoe	orchard	shrimp	waves
bud	dragonfly	hose	orchestra	shy	web
burner	drill	hydrant	owl	sieve	wheat
burrow	dusk	hygienist	paperclip	sink	whiskers
cable	earth	icicle	parents	siren	whistle
caboose	earthworm	ingredients	passageway	skates	wool
cafeteria	engineer	iron	pasture	sled	wreath
calendar	equator	judge	pedal	slingshot	x-ray
calf	eraser	junkyard	penny	sliver	
candle	excited	knight	pickax	smoke	
canoe	exhaust	ladybug	picnic	snowflake	
canyon	exhausted	lake	pill bug	soil	
cargo	fawn	landfill	planet	spatula	
carpenter	feathers	legend	plant	stall	

APPENDIX S. K-PAVE TEACHER TRAINING AGENDA

DAY 1

7:30–8:00 am	Sign-in and Introduction to Materials and Badges Continental Breakfast	
8:00–9:45 am	Welcome and Introduction Background and Orientation to the Day	
9:30–9:45 am	Go to Breakout Rooms	
9:45–11:45 am	Building Bridges Conversation Strategies: Lecture, Followed by Small Group Activities	
11:45–12:30 pm	Lunch (Practice Using *Building Strategies*)	
12:30–2:30 pm	CAR Talk Book Reading Followed by Small Group Activities	
2:30–3:30 pm	Introduce Homework Activity, Brainstorm Topics, Pick Two Books, and Put it Down in Form. (Think of a Topic or Two That has Some Good Vocabulary Potential. Consider [Bring] Possible Materials [Books, Activities, etc.] for the Next Day.)	

DAY 2

7:30–8:00 am	Continental Breakfast Questions Regarding Homework	
8:00–10:15 am	New Vehicles: Explicit Vocabulary Practices for Kindergarten Classrooms	
10:15–12:00 noon	Putting it All Together: Scheduling, Tracking, Communication with Families, Using Assistants	
12:00–12:45 pm	Lunch (Talk About *Small Group Classroom Management*)	
12:45–2:00 pm	Preparing Your First *K-PAVE* Unit, Using Your Lesson Plan, and Creating Your Own Unit Using Homework Materials	
2:00–2:45 pm	Meet with TA and Plan	
2:45–3:30 pm	Overview of the Fidelity Observation, Questions and Schedule for the Classroom Observation(s), Remediation Schedule, Schedule for Telephone Follow-ups, Schedule for Child Assessments **or for Non-Focal Teachers** Continue to Plan; Schedule for Telephone Follow-ups	

K-PAVE ASSISTANT TEACHER SCHEDULE (DAY 2 ONLY)

7:30–8:00 am	Continental Breakfast Welcome and Introduction	
	Introduction of Teacher Assistant Trainers	

8:00–9:45 am	Building Bridges: Strategies for Carrying Out Conversations with Kindergarten Children
9:45–11:45 am	CAR Talk: Book Reading Strategies
11:45–12:30 pm	Lunch (*Small Group Classroom Management)*
12:30–2:00 pm	New Vehicles: Engaging Children in the Use of Vocabulary Words
2:00–2:30 pm	Meet with Teachers to Plan

APPENDIX T. K-PAVE TEACHER PHONE FOLLOW-UP AGENDA

Attending:

Absent:

Moderator:

Assistant:

I. GENERAL PROBLEMS (ASK EACH FOLLOW-UP)

- Which is the most difficult aspect of the program for you to carry out? Why?
  - Follow-up: Does anyone have a solution to this problem they would like to share?
  - Add our own suggestions.
  - Reiterate these two questions until teachers' questions are tapped.
- Which practices do you feel will be difficult to carry out in the long term?
  - Does anyone have an idea that might be helpful to enable you to carry out this part of the program in the long term?
  - Add our own suggestions
  - Reiterate these two questions until teachers' questions are tapped.

II. PROGRAM AREAS

Building Bridges (1st follow-up)

- Is anyone having trouble with the Building Bridges aspect of the program? Can you describe the problem?
  - Follow-up: Does anyone have a solution to this problem they would like to share?
- Add our own suggestions.
- Reiterate these two questions until teachers' questions are tapped.
- Could someone share some vocabulary recasting or introducing rare words they have managed to carry out in small group conversations?

- Has expanding on children's speech started to feel natural yet? Can someone recall a particular instance where you have done this? Has anyone heard a child use this expansion in their own speech?
- Can someone share an open-ended question that seemed to yield a lot of spontaneous talk from the children?
- Can someone describe a strategy they have used to initiate child-centered conversations with children?
- Has anyone had success in scheduling the Building Bridges conversations? When are good times for scheduling this?
- Has anyone had success using the Building Bridges Tracking tool to schedule the small group conversations?

New Vehicles (2nd follow-up)

- Is anyone having problems with the NEW Vehicles aspect of the program? Can you describe the problem?
- Follow-up: Does anyone have a solution to this problem they would like to share?
    - Add our own suggestions.
    - Reiterate these two questions until teachers' questions are tapped.
- Can someone describe to me how you are using Quick Definitions in your classroom?
- Can someone describe to me how you have fit the N3C presentation of new words into your teaching?
- Can someone describe to me how you have emphasized vocabulary words so that the children understand that there is a list to be learned?
- Does anyone have a good extension activity that you are using to support vocabulary, even if it is not one that is presented in the unit we provided you? Are there any that haven't worked very well?
- Has anyone had success in scheduling the extension activities? Where works best?
- Has anyone had success keeping track of which child has spontaneously used a particular vocabulary word using the **New Vehicles Extension Activity Tracking Tool?**

CAR Talk (3rd follow-up)

- Is anyone having problems with the CAR Talk aspect of the program? Can you describe the problem?
    - Follow-up: Does anyone have a solution to this problem they would like to share?
    - Add our own suggestions.
    - Reiterate these two questions until teachers' questions are tapped.
- Has anyone tried to use the concrete questions we provided with the units? What are some of the successes or challenges that you have had creating concrete questions for your own lessons or other books you are reading to the class? Could someone share a concrete question you have used this week?
- Have you used the relational questions we have provided with the units? What are some of the successes or challenges that you have had creating relational questions for your own lessons or other books you are reading to the class? Could someone share a relational question you have used this week?
- Have you used the abstract questions we have provided with the units? What are some of the successes or challenges that you have had creating abstract questions for your own

lessons or other books you are reading to the class? Could someone share an abstract question you have used this week?

- Could someone share their difficulties or success strategies for the two whole-group book readings that you do with your class?
- Can someone share how they have scheduled three small-group sessions for each child in their classroom? Has anyone had success using the CAR Talk Tracking Tool?
- Is anyone having trouble finding new books to use in the development of their own lessons? Could someone share how they have been selecting their own books? What aspects of good books have you been searching for when you are selecting new texts to use with your children (this can be a time to remind them about depictable words, rare words etc.; the elements that we specified when we talked about the creation of the teacher's own units)?

III. WHAT CAN WE DO TO HELP YOU WITH THE PROGRAM?

IV. REMINDER OF WHEN NEXT PHONE CALL WILL BE.

APPENDIX U. SAMPLE MEANS AND STANDARD DEVIATIONS FOR STUDENT AND CLASSROOM OUTCOME MEASURES, BY INTERVENTION STATUS

Table U1. Sample intervention and control group means for student outcome measures

	K-PAVE intervention group (n = 596 students)		Control group (n = 700 students)	
Measure	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Expressive vocabulary	93.18	11.26	91.95	11.35
Listening comprehension	89.93	12.91	88.42	13.00
Academic knowledge	456.51	12.54	455.05	13.48

Table U2. Sample intervention and control group means for classroom instruction outcome measures

	K-PAVE intervention group K-PAVE(n = 60 classrooms)		Control group (n = 68 classrooms)	
Measure	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Vocabulary and comprehension support	0.44	0.94	–0.41	0.88
Instructional support	3.66	1.49	2.93	1.28
Emotional support	5.49	0.95	5.25	0.80
Time spent on literacy in areas other than vocabulary and comprehension	0.52	0.27	0.50	0.33

APPENDIX V. CHECKING MODEL ASSUMPTIONS

MODEL EXAMINING IMPACTS ON STUDENTS' EXPRESSIVE VOCABULARY (EVT–2)

The model in Appendix K assumes error terms for the student-, classroom-, and school-level equations that are normally distributed with mean 0. The tests of the statistical significance of estimated treatment impacts depend on this property. So testing the normality assumption by looking at the residuals from the fitted equations is worthwhile.

Normality assumption

Whether the assumption of normally distributed errors had been met was evaluated by examining the distribution of studentized residuals at the student level and by examining residuals at each level plotted against a normal distribution. Examination of the residuals indicates that the normality assumption was met. The shape of the distribution of studentized residuals appears normal, although its tails are long, with values ranging from –6.6 to 6.9. Even with some values that appeared extreme, 4.3% of residuals were more than two standard deviations from the mean, which is consistent with a normal distribution. The plot of the studentized residuals at the student level versus a normal distribution shows a straight line with values at the tails deviating from the line somewhat (see left panel of Figure V1). The plot remains straight with the removal of six potential outliers, which have studentized residuals from the student-level equation that are four or more standard deviations from the mean (see right panel of Figure V1).

Findings from the impact model remained consistent when these six potential outliers were excluded from the analysis (see sensitivity analyses in Appendix N).

Figure V1. Studentized residuals at the student level plotted against a normal distribution

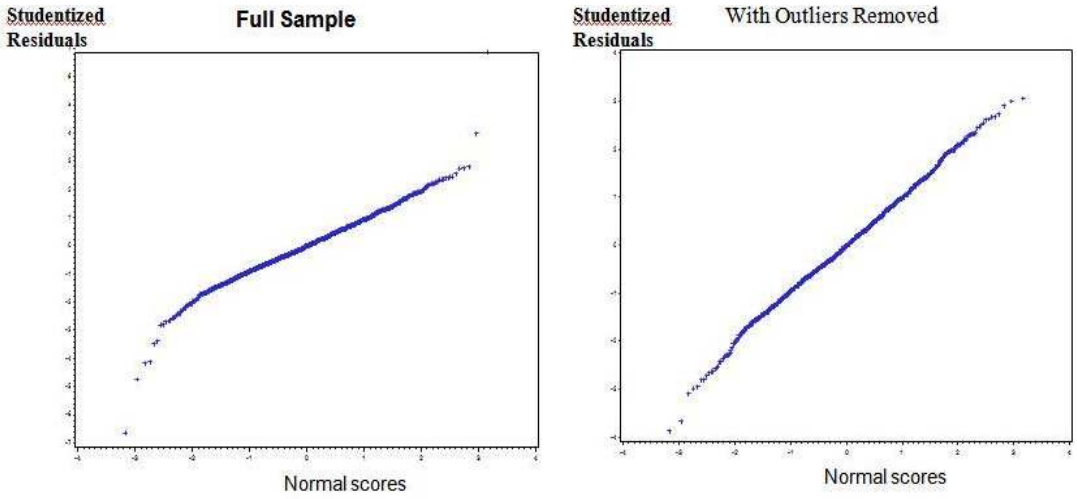
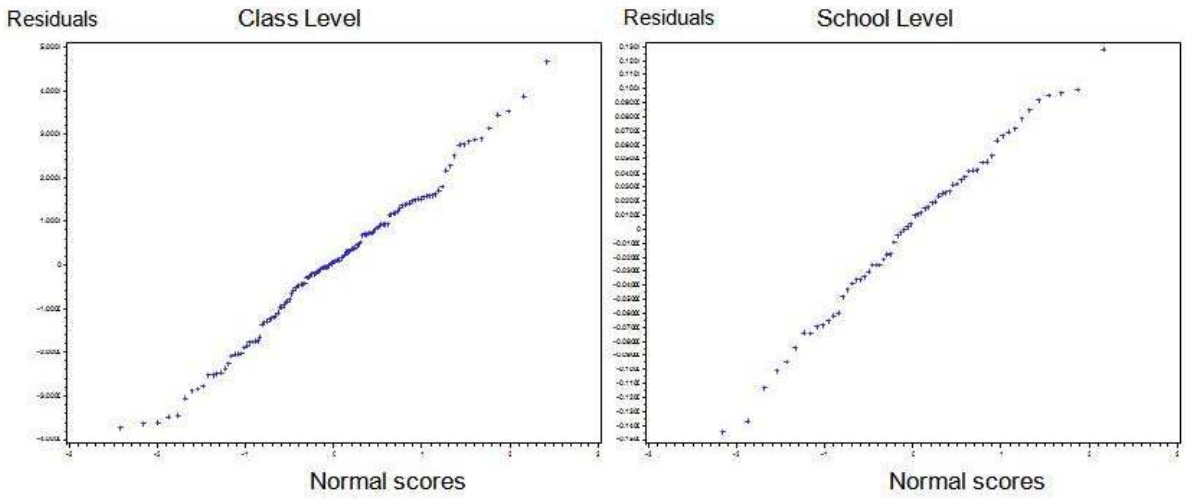


Figure V2 shows a plot of raw residuals versus a normal distribution at the classroom level (left panel) and at the school level (right panel). In both plots, the residuals fall approximately on a straight line indicating that they are approximately normally distributed.

Figure V2. Raw residuals versus normal distribution at classroom and school levels



Homoscedasticity assumption

Whether the homoscedasticity assumptions in the hierarchical model were met was evaluated by examining plots of residuals versus the outcome (expanded vocabulary test [EVT] posttest) at the student, classroom, and school levels, for both treatment and control classrooms.

Tests of the statistical significance of the impact estimate depend on meeting these assumptions. Heteroscedasticity (the lack of homoscedasticity) would appear as wider dispersion of studentized residuals at some values of the EVT posttest outcome measure than at other values. The plots indicate that the homoscedasticity assumption was met (Figure V3). The full plot (left panel of Figure V3) suggests that there are four potential outliers in the lower range of EVT posttest scores that are spread apart from the point cloud. The plot without these four cases (right panel of Figure V3) shows an even distribution of residuals across the range of EVT posttest scores. Figure V4 presents plots of the residuals for the classroom level (left panel) and school level (right panel), which also show equal variation across the range of EVT posttest scores.

Figure V3. Studentized residuals at the student level plotted against EVT posttest score for students in treatment (red) and control group (blue) schools

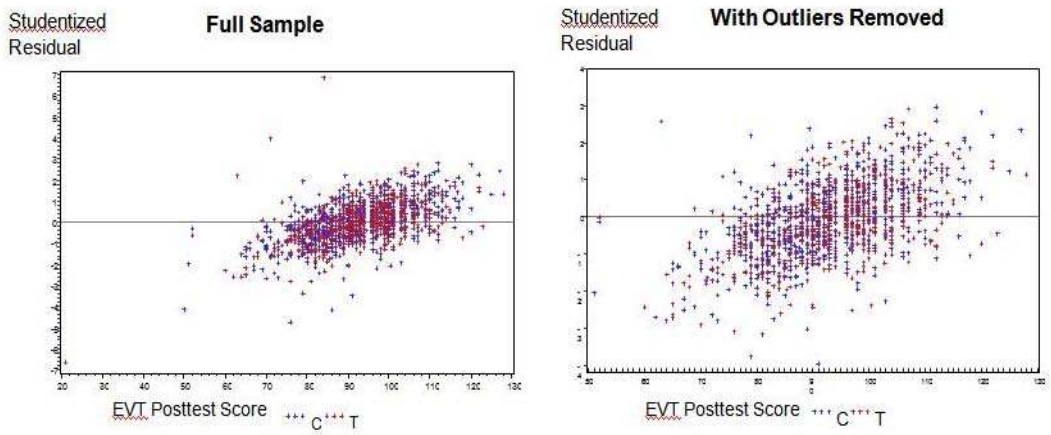
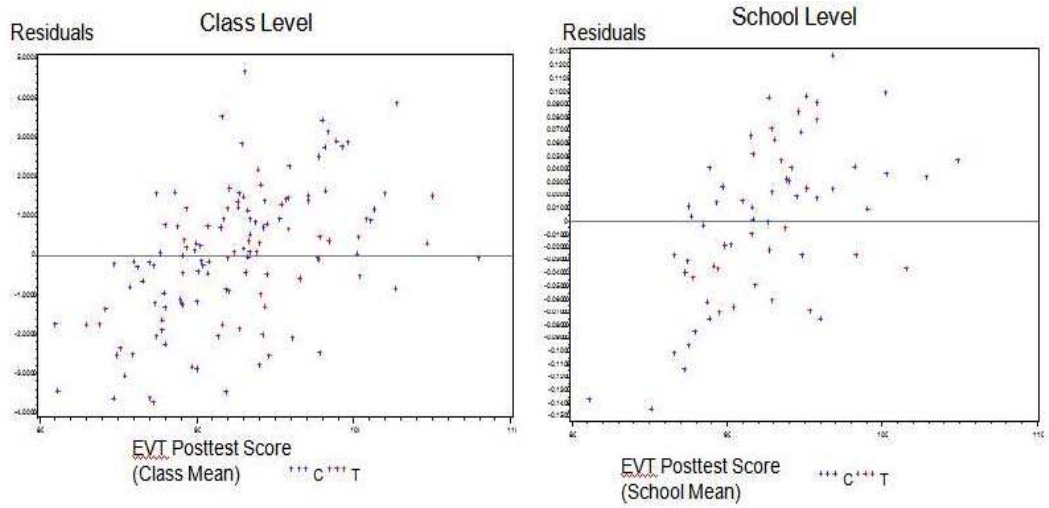


Figure V4. Studentized residuals at the classroom level (left panel) and school level (right panel) plotted versus EVT posttest score for students in treatment (red) and control group (blue) schools

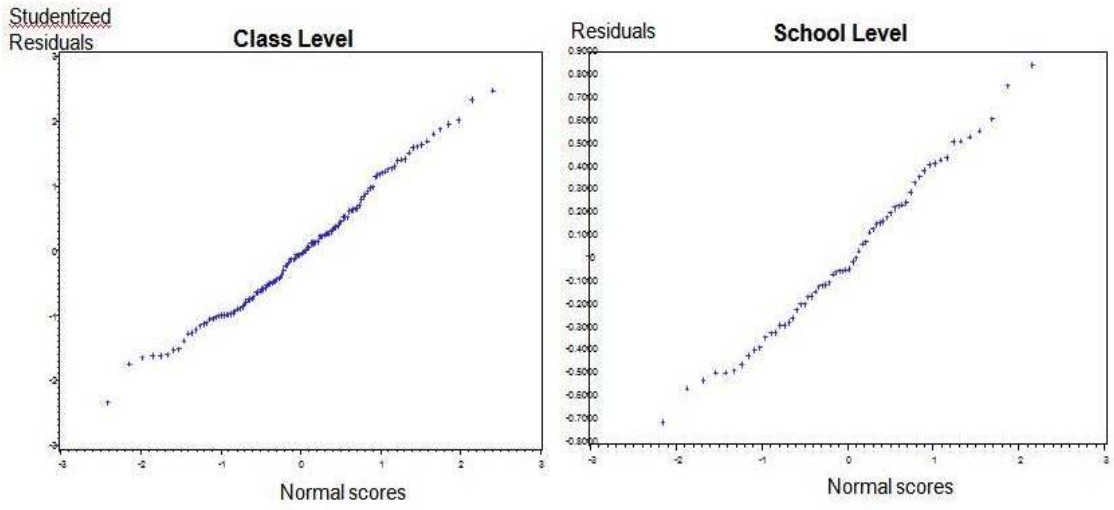


Vocabulary and comprehension support

Normality assumption

The residuals at each level versus a normal distribution were plotted to evaluate whether the assumption of normally distributed errors had been met (Figure V5). At the classroom level (left panel) and the school level (right panel), the residuals fall approximately along a straight line, indicating that they are approximately normally distributed. The shape of the distribution of studentized residuals at the classroom level appears normal, with no indication of extreme values. Studentized residuals ranged from –2.3 to 2.5, and 3% of the residuals had values more than two standard deviations from the mean. There was no indication of any outliers.

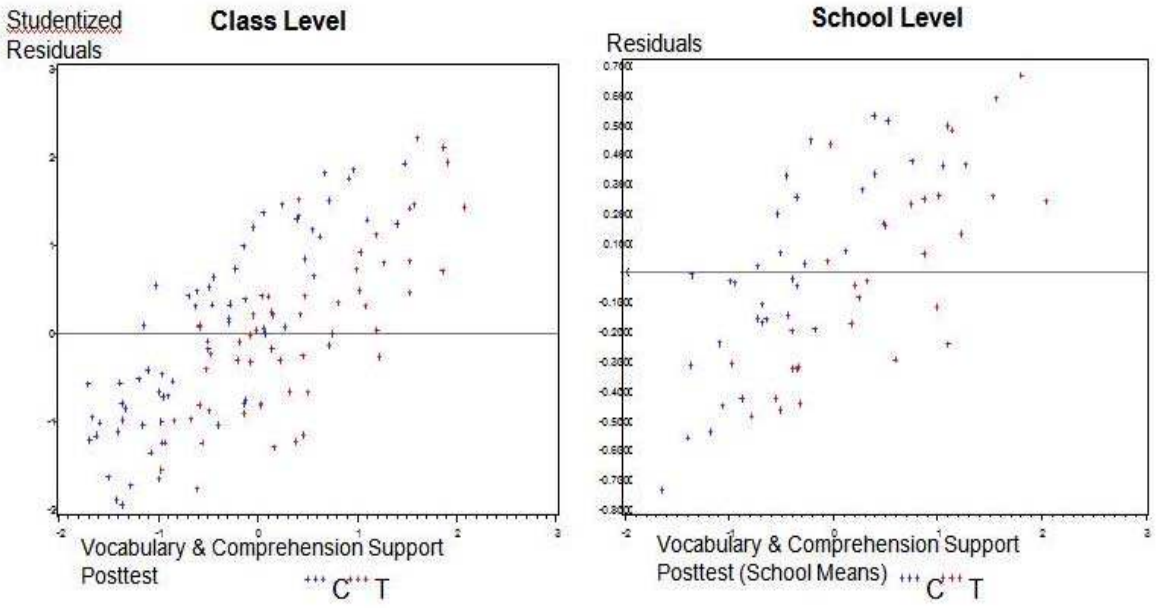
Figure V5. Plot of residuals versus a normal distribution at the classroom and school levels, from the model testing the impact on vocabulary and comprehension support



Homoscedasticity assumption

Plots of residuals versus the outcome (vocabulary and comprehension support posttest) at the classroom and school levels were examined to evaluate whether the homoscedasticity assumption was met (Figure V6). Both plots indicate that the homoscedasticity assumption was met. Residual variation is equal across the range of posttest scores, and no curves or funnel shapes are apparent.

Figure V6. Residuals at the classroom level and school level versus vocabulary and comprehension support posttest score for students in treatment (red) and control group (blue) schools

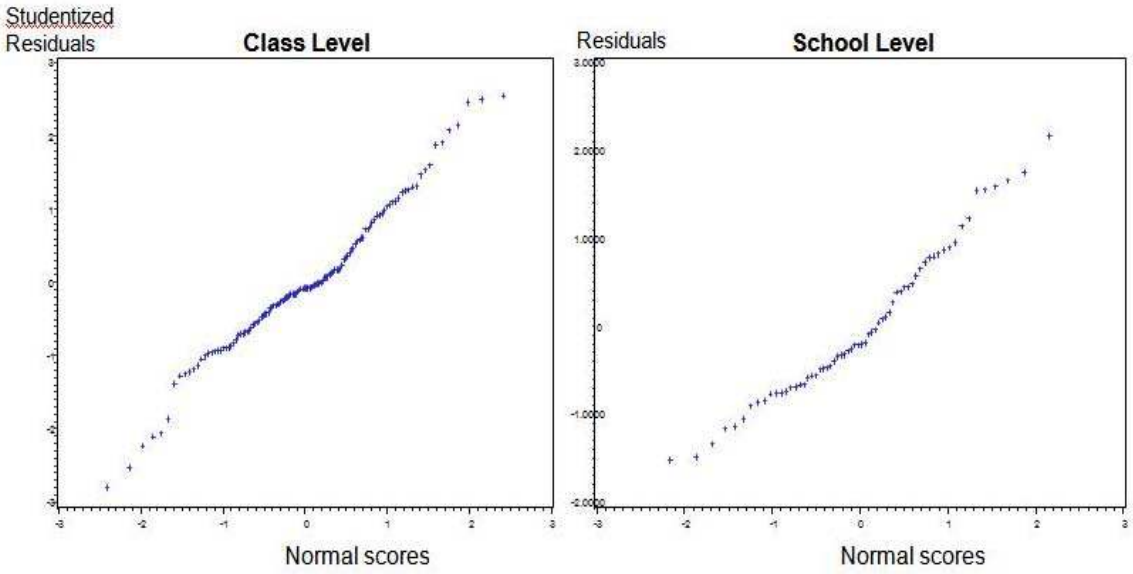


Instructional support

Normality assumption

The normality assumption was evaluated by examining residuals at each level plotted against a normal distribution (Figure V7). At the classroom level (left panel) and the school level (right panel), the residuals fall along a straight line, indicating that they are normally distributed. No data points in the tails warrant further investigation. The shape of the distribution of studentized residuals at the classroom level appears normal. However, 7.8% of residuals have values greater than two standard deviations from the mean, which is above the 5% expected in a normal distribution. Studentized residuals ranged from –2.8 to 2.6. There was no indication of any outliers.

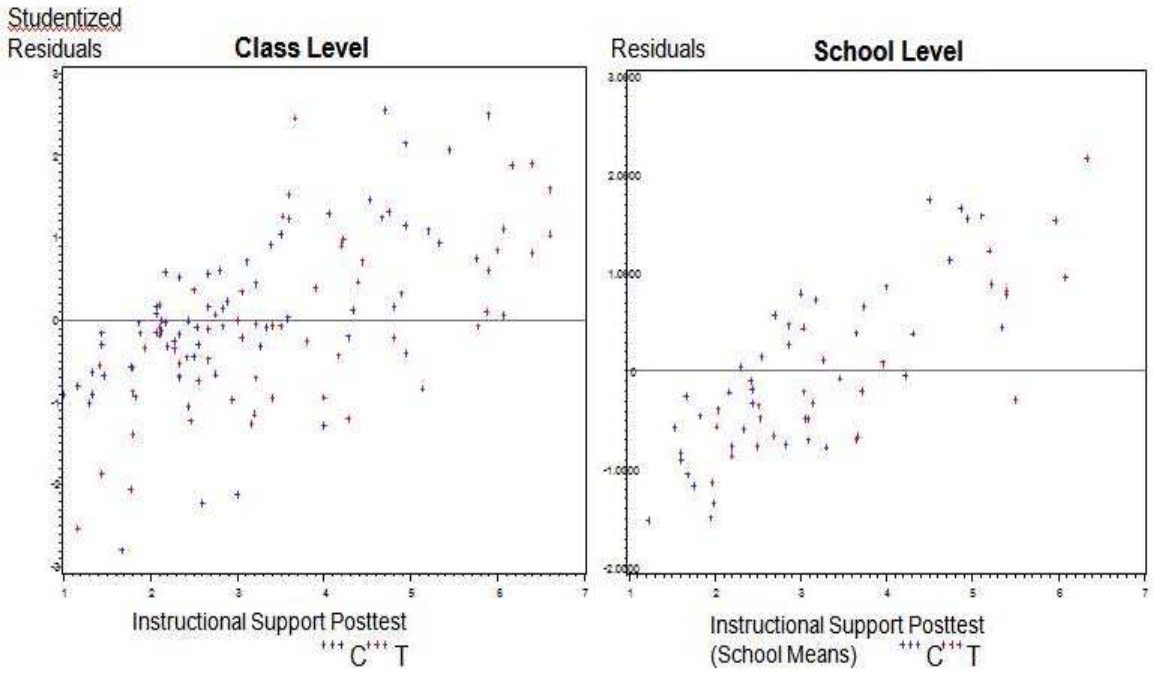
Figure V7. Plot of residuals versus a normal distribution at the classroom and school levels, from the model testing the impact on instructional support



Homoscedasticity assumption

Plots of classroom- and school-level residuals versus the outcome (instructional support posttest) were examined to evaluate whether the homoscedasticity assumption was met (Figure V8). Both plots indicate that the homoscedasticity assumption was met. Residual variation is equal across the range of posttest scores, and no curves or funnel shapes are apparent.

Figure V8. Residuals at the classroom level and school level versus instructional support posttest score for students in treatment (red) and control group (blue) schools



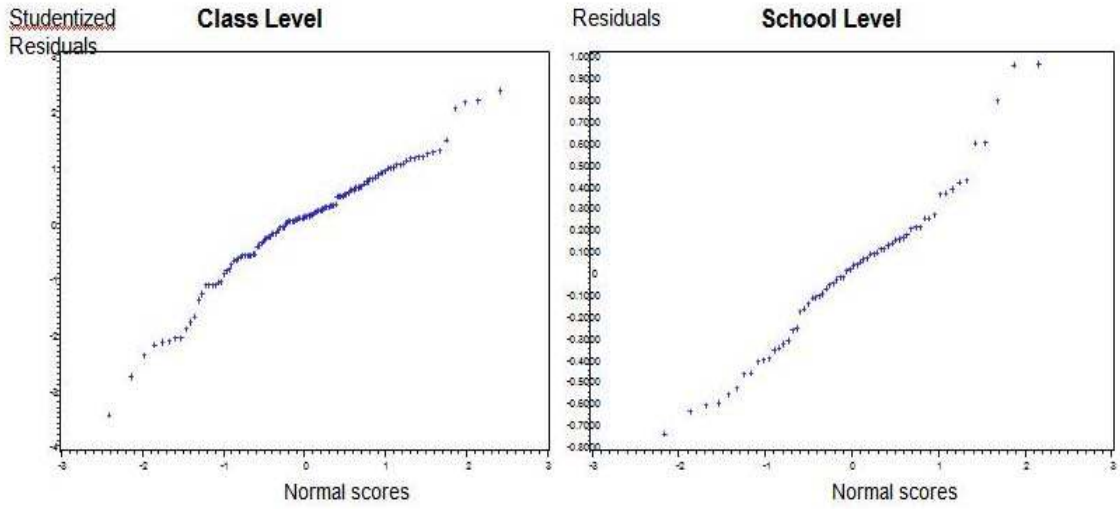
Emotional support

Normality assumption

The normality assumption was evaluated by examining residuals at each level plotted against a normal distribution (Figure V9). At the classroom level (left panel) and the school level (right panel), most of the residuals fall along a straight line; however, data points in both plots show some curve in the tails. Although the plots suggest that the normality assumption was met, the data points in the tail were investigated further. Studentized residuals ranged from –3.5 to 2.4. The shape of the distribution of residuals was normal. There were 9.4% of residuals more than two standard deviations from the mean, which is slightly more than the 5% expected for a normal distribution.

Based on an examination of the distribution of the studentized residuals, one potential outlier was identified, with a studentized residual of –3.45. Findings from the impact model remained consistent when this potential outlier was excluded from the analysis (see sensitivity analyses in Appendix N.)

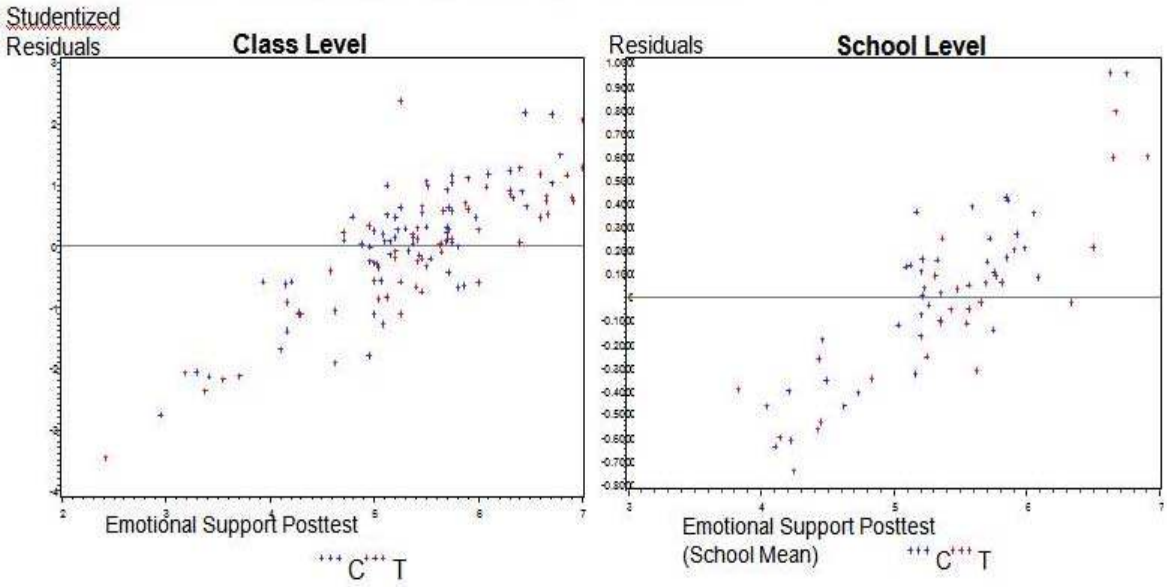
Figure V9. Plot of residuals versus a normal distribution at the classroom and school levels, from the model testing the impact on emotional support



Homoscedasticity assumption

Plots of classroom- and school-level residuals versus the outcome (emotional support posttest) were examined to evaluate whether the homoscedasticity assumption was met (Figure E10). Both plots indicate that the homoscedasticity assumption has been met. Residual variation is equal across the range of posttest scores, and no curves or funnel shapes are apparent.

Figure V10. Residuals at the classroom level and school level versus emotional support posttest score for students in treatment (red) and control group (blue) schools



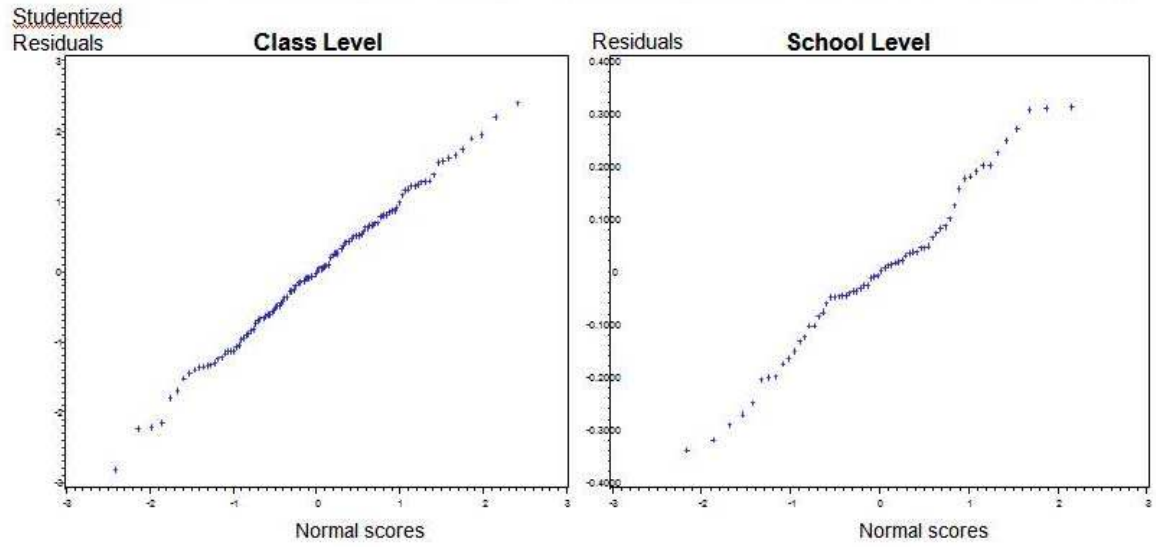
PROPORTION OF CLASSROOM OBSERVATION CYCLES SPENT ON NONVOCABULARY LITERACY INSTRUCTION

Normality assumption

The normality assumption was evaluated by examining residuals at each level plotted against a normal distribution (Figure V11). At the classroom level (left panel) the residuals fall along a straight line. At the school level (right panel), some curve is apparent in the middle and the upper end of the line. Even with the curve at the school level, the assumption of normality appears to have been met. The shape of distribution of the classroom level residuals appears normal. The studentized residuals range from –2.8 to 2.4, and 4.7% of classrooms have residuals that are more than two standard deviations from the mean, which is consistent with a normal distribution.

Examination of the distribution of the studentized residuals did not indicate that there were any extreme values. However, because some curve was observed in the extremes of the normal probability plot (see Figure V11), the six classrooms with studentized residuals greater than +/– 2.0 from the impact model were removed as a sensitivity analysis (see Appendix N). Due to the slight curviness at the center of the plot of the school level residuals, a Wilks-Shapiro test was conducted, and the null hypothesis that the residuals were not normally distributed ($W = 0.99$, $p = .70$) could not be rejected. Findings remained consistent when the six observations were excluded.

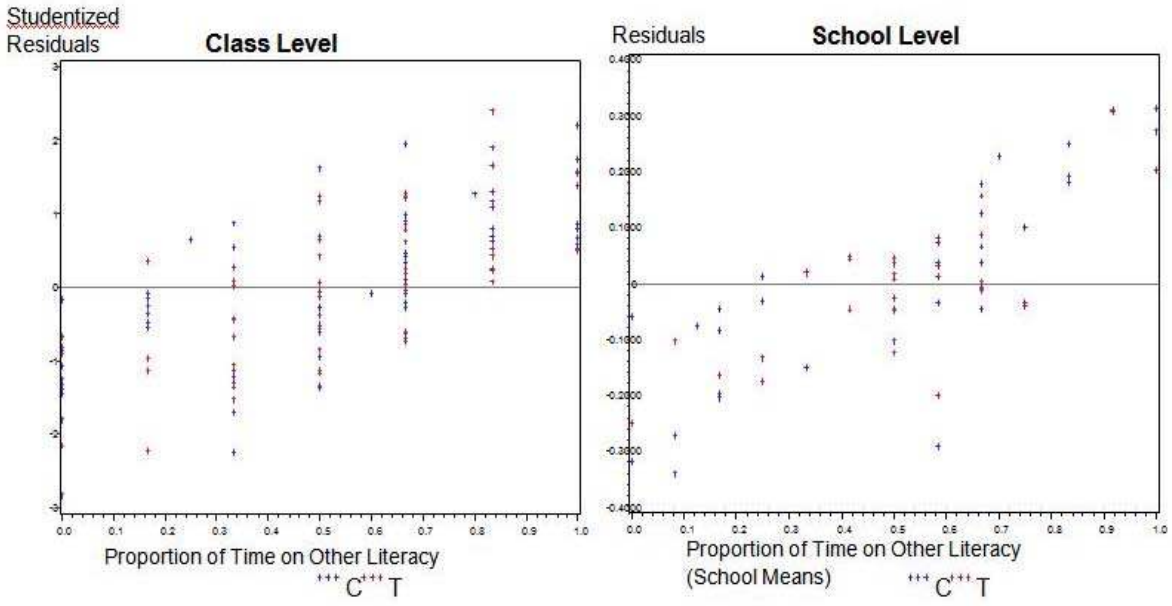
Figure V11. Plot of residuals versus a normal distribution at the classroom and school levels, from the model testing the impact on nonvocabulary literacy instruction



Homoscedasticity assumption

Plots of classroom- and school-level residuals versus the outcome were examined to evaluate whether the homoscedasticity assumption was met (Figure V12). Both plots indicate that the homoscedasticity assumption was met. Residual variation is equal across the range of posttest scores, and no curves or funnel shapes are apparent.

Figure V12. Residuals at the classroom level and school level versus nonvocabulary literacy instruction for students in treatment (red) and control group (blue) schools



APPENDIX W. TRANSLATING IMPACTS ON STUDENTS INTO AGE-EQUIVALENT DIFFERENCES IN POSTTEST OUTCOMES

To interpret the practical meaning of the magnitude of the impact of K-PAVE on students' expressive vocabulary and academic knowledge, posttest differences in scale scores between the intervention and control group were translated into age-equivalent differences. Ideally, regression-adjusted posttest means for both groups could be translated into age-equivalent scores, but the scale scores used in this particular analysis cannot be directly translated into age-equivalent scores. They must be translated into raw scores (the number of correct items) for each test, and the raw scores then converted into age-equivalent scores. Raw score conversion into age equivalents is possible because, based on the norming sample for each test, the average raw score was determined for students at each month of age. Hence, each raw score has an age equivalent.

Raw scores were not analyzed in the test for impacts. For the Expressive Vocabulary Test–2nd Edition (EVT–2), standard scores were analyzed—that is raw scores transformed (for students at each age in months) into a scale with a mean of 100 and a standard deviation of 15, based on the norming sample. For academic knowledge, W-scores, which are item response theory–scaled scores, were reviewed.

The two-stage process of converting the regression-adjusted mean posttest scores first into raw scores and then into age-equivalent scores is not precise. The statistical precision advantage of basing the difference on regression-adjusted scores may be lost in the process. For this reason, two approaches to generating age-equivalent scores were examined: converting regression-adjusted mean scores into raw scores and then using those raw scores to generate age-equivalent scores and using sample mean raw scores for each group without regression adjustment to generate age-equivalent scores. For both tests, both approaches yield consistent results.

For the EVT–2, the standard-to-raw-score conversion table for the spring of kindergarten in the EVT–2 manual (Williams 2007, p. 178) was used to convert regression-adjusted mean standard scores to raw scores. The regression-adjusted mean standard scores of 93 for the intervention group and 92 for the control group can be converted to raw scores of 69 for the intervention group and 67–68 for the control group for test form A. These regression-adjusted raw scores closely approximate the observed posttest raw score means of 69.3 (for the intervention group) and 68.0 (for the control group). Using the raw score to age-equivalent score conversion table in the EVT–2 manual (pp. 192–93), both the regression-adjusted mean posttest score and the sample mean raw score in the intervention group are equivalent to the average score for a child who is 5 years, 7 months, and the regression-adjusted mean posttest score and the sample mean raw score in the control group are equivalent to the average score for a child who is 5 years, 6 months. Hence, the regression-adjusted mean difference is one month.

For the Woodcock-Johnson III/ Normative Update academic knowledge test, W-scores map closely to raw scores but do not match exactly. For example, in the current sample, students with raw scores of 30 have W-scores ranging from 454 to 456, and students with raw scores of 31 have W-scores ranging from 456 to 459. The age-equivalent score for a raw score of 30 is 5 years, 9 months, and the age equivalent score for a raw score of 31 is 5 years, 11 months. If the

regression-adjusted mean scores for the intervention and control groups are converted to raw scores and then to age-equivalent scores, the age-equivalent score for the control group is 5 years, 9 months, and the age-equivalent score for the intervention group is between 5 years, 9 months and 5 years, 11 months—or approximately 5 years, 10 months. Alternatively, each individual student's raw score can be converted into its age-equivalent score and the age-equivalent scores for the intervention and control groups averaged. This approach—although not regression-adjusted—yields a similar result: the average age-equivalent score is 5 years, 9.3 months for the control group and 5 years, 10.5 months for the intervention group. Using either approach, the intervention group is ahead of the control group in academic knowledge by an average of one month.

REFERENCES

Adams, M. J., Foorman, B. R., Lundberg, I., & Beeler, T. (1998). *Phonemic awareness in young children: A classroom curriculum*. Baltimore: Brookes Publishing.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Washington, DC: National Academy of Education, Commission on Education and Public Policy.

Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, J. Jensen, D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 752–785). New York: Macmillan.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *Elementary School Journal, 107*(3), 251–271.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford.

Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *American Educator*, *25*(1), 24–29.

Biemiller, A., & Boote, C. (2006). An effective method for building vocabulary in primary grades. *Journal of Educational Psychology, 98*(1), 44–62.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93*(3), 498–520.

Bishaw, A., & Iceland, J. (2003). *Poverty: 1999. Census 2000 Brief.* Washington, DC: U.S. Census Bureau.

Bowman, B., Donovan, M., & Burns, M. (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

Brabham, E. G., & Lynch-Brown, C. (2002). Effects of teachers' reading-aloud styles on vocabulary acquisition and comprehension of students in the early elementary grades. *Journal of Educational Psychology, 94*(3), 465–473.

Burghardt, J., Deke, J., Kisker, E., Puma, M., & Schochet, P. (2009). *Regional educational laboratory rigorous applied research studies: Frequently asked analysis questions.* Institute

of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematics Policy Research.

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.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language (CASL).* Bloomington, MN: Pearson Assessments.

Catts, H. W., Hogan, T. P., & Adolf, S. M. (2005). Developmental changes in reading and reading disabilities. In H.W. Catts & A.G. Kahmi (Eds.), *The connections between language and reading disabilities* (pp. 25–40). Mahwah, NJ: Lawrence Erlbaum Associates.

Chall, J. S., and Conard, S. S. (1991). *Should textbooks challenge students?* New York: Teachers College Press.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disabilities Quarterly*, *30*, 74–88.

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth versus depth. *The Elementary School Journal, 110*(1), 1–18.

Coyne, M. D., Simmons, D. C., Kame'enui, E. J., & Stoolmiller, M. (2004). Teaching vocabulary during shared storybook readings: An examination of differential effects. *Exceptionality, 12*(3), 145–162.

Curtis, M. E. (1987). Vocabulary testing and instruction. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 37–51). Hillsdale, NJ: Erlbaum.

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

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24*(1), 174–186.

Fishman, G. S., and Moore, L. R. (1982). A statistical evaluation of multiplicative congruential random number generators with modulus 2^{31}. *Journal of the American Statistical Association, 77*(377), 129–136.

Forder, P. M., Gebski, V. J., and Keech, A. C (2005). Allocation concealment and binding. *The Medical Journal of Australia, 182*(2), 87–89.

Goodson, B. D., Layzer, C. J., Smith, W. C., & Rimdzius, T. (2004). *Observation Measures of Language and Literacy Instruction (OMLIT)*. Unpublished instrument. Cambridge, MA: Abt Associates.

Greenberg, M. T., Domitrovich, C. E., & Bumbarger, B. (2001). The prevention of mental disorders in school-aged children: Current state of the field. *Prevention and Treatment, 4*(1), 1–64.

Hamilton, C. E., & Schwanenflugel, P. J. (under contract). *PAVEd for success: Supporting vocabulary and oral language development in prekindergarten and kindergarten children*. Baltimore: Brookes.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–83). Baltimore, MD: Brookes Publishing.

Hargrave, A. C., & Senechal, M. (2000). Book reading intervention with language-delayed preschool children: The benefits of regular reading and dialogic reading. *Journal of Child Language, 15*, 765–790.

Hart, B., & Risley, R. T. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Brookes Publishing.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*, 303–317.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172-177.

Houghton Mifflin Reading. (2008). Orlando, FL: Houghton Mifflin Harcourt School Publishers.

Howes, C., Burchinal, M., Pianta, R. C., Bryant, D., Early, D., & Clifford, R. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27–50.

Jones, M. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91,* 222-230.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement, Second Edition (KTEA-II)*. Shoreview, MN: AGS Publishing.

Landry, S. H., Anthony, J. L., Swank, P., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology, 101*(2), 448-465.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Technical Manual. Woodcock-Johnson III Normative Update.* Rolling Meadows, IL: Riverside Publishing.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge, UK: Cambridge University Press.

Morrison, F., & Connor, C. M. (2002). Understanding schooling effects on early literacy. *Journal of School Psychology, 40*(6), 493–500.

National Early Literacy Panel. (2009). *Developing early literacy: Report of the National Early Literacy Panel, A scientific synthesis of early literacy development and implications for intervention.* Washington, DC: National Institute for Literacy, National Center for Family Literacy.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development. ERIC Document Reproduction Service No. ED444127)

Penno, J. F., Wilkinson, A. G., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect*? Journal of Educational Psychology, 94*(1), 23–33.

Pianta, R. C. (2006). Teacher-child relationships and early literacy. In D. Dickinson & S. Neuman (Eds.), *Handbook of early literacy research, Vol. 2* (pp. 149–162). New York: The Guilford Press.

Pianta, R. C., LaParo, K., & Hamre, B. (2008). *Classroom Assessment Scoring System (CLASS K–3)*. Baltimore, MD: Brookes Publishing.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center

for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Quality Counts. 2008. *Special supplement. Mississippi state highlights* (p. 5). Bethesda, MD: Editorial Projects in Education Research Center.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods. 2nd edition*. Newbury Park, CA: Sage.

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, Speech and Hearing Services in the Schools, 37*(1), 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.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology, 40*(6), 451–475.

Schochet, P. (2008a) Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87.

Schochet, P. (2008b). *Guidelines for multiple testing in impact evaluations. Technical methods report* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved on October 20, 2009, from http://ies.ed.gov/ncee/pdf/20084018.pdf

Schochet, P. Z. (October, 2009). *Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes?* (NCEE 2009-4065). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved on December 1, 2009, from http://ies.ed.gov/ncee/pdf/20094065.pdf

Schwanenflugel, P. J., Hamilton, C. E., Bradley, B. A., Ruston, H. P., Neuharth-Pritchett, S., & Restrepo, M. A. (2005). Classroom practices for vocabulary enhancement in prekindergarten: Lessons from PAVEd for Success. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 155–178). Mahwah, NJ: Lawrence Erlbaum Associates.

Schwanenflugel, P. J., Hamilton, C. E., Neuharth-Pritchett, S., Restrepo, M. A., Bradley, B. A., & Webb, M. Y. (In press). PAVEd for Success: An evaluation of a comprehensive literacy program for 4-year-old children. *Journal of Literacy Research*.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4).* Bloomington, MN: Pearson Assessments.

Senechal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language, 24*(1), 123–128.

Senechal, M., & Cornell, E. H. (1993). Vocabulary acquisition through shared reading experiences. *Reading Research Quarterly, 28*(4), 360–374.

Senechal, M., Thomas, E., & Monker, J. (1995). Individual differences in 4-year-old children's acquisition of vocabulary during storybook reading. *Journal of Educational Psychology, 87*(2), 218–229.

SERVE Center, University of North Carolina at Greensboro & University of Georgia (June, 2008). *Kindergarten PAVEd for Success (K-PAVE): A Program to Enhance Kindergarteners' Vocabulary, Teacher's Guide.* Greensboro, NC: The SERVE Center.

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.

Smith, W. C., Dwyer, M. C., Dixon, Q., Schimmenti, J., Boulay, B., Khalil, B., Blocklin, M., & Gamse, B. (2005). *The instructional practice in reading inventory (IPRI)*. Unpublished instrument. Cambridge, MA: Abt Associates.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: The effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology, 41*(1), 225–234.

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.

Strickland, D. S., Danske, K., & Monroe, J. K. (2002). *Supporting struggling readers and writers: Strategies for classroom intervention 3–6.* Portland, ME: Stenhouse Publishers.

Treasures: A reading/language-arts program. (2008). New York: MacMillan/McGraw-Hill.

Trophies. (2005). Orlando, FL: Houghton Mifflin Harcourt School Publishers.

U.S. Census Bureau. (2008). Poverty. Retrieved on October 28, 2010, from http://www.census.gov/hhes/www/poverty/ poverty.html

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. Scientific Studies of Reading, 11(1), 3–32.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/ L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217-254.

Wasik, B. A., & Bond, M. A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology, 93*(2), 243–250.

Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63-74.

Wendling, B. J., Schrank, F. A., & Schmitt, A. J. (2007). *Educational interventions related to the Woodcock-Johnson III Tests of Achievement* (Assessment Service Bulletin No. 8). Rolling Meadows, IL: Riverside Publishing.

White, T. G., Graves, M. F., & Slater, W. H. (1990). Growth of reading vocabulary in diverse elementary schools: Decoding and word meaning. *Journal of Educational Psychology, 82*(2), 281–290.

Williams, K. T. (2007). *Expressive Vocabulary Test–2nd Edition*. Circle Pines, NM: American Guidance Service.

Williams, K. T. (2001). *Group Reading Assessment and Diagnostic Evaluation (GRADE).* Circle Pines, NM: AGS Publishing.

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock-Johnson III Normative Update.* Rolling Meadows, IL: Riverside Publishing.

Zareva, A., Schwanenflugel, P. J., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*, 567-595.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA