NCEE 2012-4009

U. S. DEPARTMENT OF EDUCATION

Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): First Grade Follow-up Impact Report and Exploratory Analyses of Kindergarten Impacts

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

Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): First Grade Follow-up Impact Report and Exploratory Analyses of Kindergarten Impacts

Final Report

December 2011

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, Research Management Leader
Regional Education Laboratory Southeast

Project Officer:
Sandra Garcia
Institute of Education Sciences

NCEE 2012–4009
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

December 2011

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED–06-CO–0028.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

While permission to reprint this publication is not necessary, the citation should read: Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2011). *Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB): First Grade Follow-up Impact Report and Exploratory Analyses of Kindergarten Impacts* (NCEE 2012–4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the Institute of Education Sciences website at http://ncee.ed.gov and the Regional Educational Laboratory Program website at http://edlabs.ed.gov.

Alternate Formats. Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department’s Alternate Format Center at 202-260-9895 or 202-205-8113.

DISCLOSURE OF POTENTIAL CONFLICT OF INTEREST[1]

None of the authors or other staff involved in the study from the Regional Educational Laboratory Southeast at the SERVE Center at the University of North Carolina at Greensboro, Abt Associates Inc., ANALYTICA, or Empirical Education Inc. has financial interests that could be affected by the content of this report.

[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.

ACKNOWLEDGEMENTS

Many individuals and organizations contributed to the evaluation of the K-PAVE vocabulary intervention. We want to thank Dr. Hank Bounds, former State Superintendent of Education for Mississippi, and the other staff in the Mississippi Department of Education for their support in allowing us to conduct a random assignment evaluation and for communicating with district superintendents and school principals to facilitate recruitment of districts and schools for the study. In addition, we greatly appreciate the many school district superintendents, school principals, and teachers who were willing to open their classrooms and share their time with us.

Many people from the SERVE Center at the University of North Carolina at Greensboro, the prime contractor for the Regional Educational Laboratory-Southeast that funded the study, were critical to the success of the project. Frankie Walton White, Alexa Edwards, Melissa Williams, Fain Barker, Lynn Amwake, Lucy Wynn, De Fuller, Glyn Brown, Teri White, Beth Thrift, and Teresa Lewis recruited participants, coordinated the intervention training and materials, trained data collectors, collected data, and identified first grade students for the follow up study. David Schaefer was critical in the document review and editing process.

In addition, we want to thank the staff from Empirical Education Inc. for their accurate database entry of student assessments and classroom observations.

Staff from Abt Associates was invaluable to many aspects of the project, including training data collectors and assisting with impact analyses. Carter Epstein developed rigorous training sessions for child assessors and classroom observers, and she lead a team of excellent trainers, including Kenyon Maree, Rachel McCormick, Cay Bradley, Alyssa Rulf Fountain, Megan Tiano, Laurie Bozzi, and Rachel Luck. Carolyn Layzer assisted with pilot testing of data collection instruments. Amanda Parsad was a crucial member of the impact analysis team. Thanks also go to Amanda Parsad, David Cook, and Tom McCall for their assistance in preparing the restricted-use data release.

Contents

FIGURES

TABLES

Map

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. 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 the 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 foundational skills in young students that might improve reading achievement. The Mississippi State Department of Education has focused specifically on interventions that may enhance students' vocabulary knowledge. The Mississippi state legislature placed a high priority on meeting the early education needs of students in and near the Mississippi 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, this study tested the impact of a vocabulary instruction program on students' expressive vocabulary when used by kindergarten teachers in the Mississippi Delta area as a supplement to their regular instructional program. The study examined whether the intervention had impacts on students at the end of kindergarten and whether the impacts were sustained in grade 1, in the absence of additional intervention. Previous research on the program showed that a preschool version of the curriculum was associated with greater student vocabulary acquisition but did not use methods that could establish causal relationships.

Kindergarten PAVEd for Success (K-PAVE) was selected to be tested in Mississippi for three reasons. First, it is one of only a small number of vocabulary interventions are appropriate for this age group. Second, 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 addressed limitations of earlier research and ensured a valid basis for estimating both the immediate effect of K-PAVE, implemented across a range of settings in the real world, on vocabulary knowledge of students in kindergarten and any sustained effects at the end of grade 1.

K-PAVE consists of 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 book 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 24-week supplement to the regular kindergarten 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.

Results from a study of the impacts of K-PAVE on students at the end of the kindergarten intervention year were reported in a 2010 study (Goodson, Wolf, Bell, Turner, and Finney 2010). The kindergarten study of K-PAVE focused on 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, and the impact was statistically significant (the 95% confidence interval around the impact estimate was 0.4–2.8 points).The standardized effect size for this impact was 0.14. Translating this effect size into difference in age–equivalent scores, at the end of kindergarten, students who received the K-PAVE intervention were one month ahead of students in the control group in vocabulary development (see Goodson et al. 2010 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 on academic knowledge was statistically significant, with a magnitude of 1.95 points 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 95% confidence interval around the impact estimate was 0.2–3.7 points). The standardized effect size for this impact was 0.14. Translating this effect size into a difference in age–equivalent scores, at the end of kindergarten, students who received the K-PAVE intervention were one month ahead of students in the control group in academic knowledge. K-PAVE did not have a statistically detectable impact on listening comprehension.

The grade 1 follow-up study examined whether the effects of K-PAVE in kindergarten provided students in intervention schools with a sustained advantage in vocabulary and in related literacy skills in grade 1, when formal reading instruction typically begins. The study did not find any statistically significant impacts of K-PAVE at the end of grade 1 on expressive vocabulary, academic knowledge, or passage comprehension.

The follow-up study also explored differences in the impact of K-PAVE on subgroups of students and schools in both kindergarten and grade 1. Given gender differences in literacy skills at kindergarten entry (e.g., Ready, LoGerfo, Lee, and Burkam 2005), the study explored whether K-PAVE impacts differed for girls and boys. It did not find a statistically significant difference in the estimated impact of K-PAVE on girls and boys on any student outcomes measured in kindergarten or grade 1. The null hypothesis that the impact of K-PAVE is the same for both girls and boys could therefore not be rejected for any of the student outcome measures.

The study also examined whether impacts of K-PAVE differed for students entering kindergarten at different levels of expressive vocabulary, academic knowledge, and comprehension. Students with pretest scores at least one standard deviation below the age-normed mean were considered to have "low" pretest scores, and those scoring above that threshold were considered to have pretest scores that were "not low." Students with low pretest scores may not be as prepared to benefit from K-PAVE as their higher-scoring peers; alternatively, K-PAVE may offer students with low pretest scores an opportunity to catch up with their higher scoring peers. However, no statistically significant difference was found in the average impact of K-PAVE on students with and without low pretest scores on any student outcomes measured in kindergarten or grade 1. Therefore, the null hypothesis that the impact of K-PAVE is the same for students with and without low pretest scores could not be rejected.

Differences in impacts of K-PAVE for Reading First and non-Reading First schools were also explored because Reading First schools may have already been using high-quality literacy instruction practices in kindergarten, and the addition of K-PAVE may have made less of a difference than in non-Reading First schools. Alternatively, teachers in Reading First schools may have a deeper understanding of children's literacy and have been better able to implement K-PAVE, which could lead to larger impacts of K-PAVE in Reading First schools. There was a statistically significant difference in the impact of K-PAVE between Reading First and non-Reading First schools on one outcome—academic knowledge measured at the end of kindergarten. The average impact of K-PAVE for students in Reading First schools was 3.8 points lower than for students in non-Reading First schools (t = 3.08, p = .002), a difference of 0.28 standard deviation (the 95% confidence interval around the impact estimate was –7.6 to –0.04 points). In non-Reading First schools, there was a positive and statistically significant impact of K-PAVE on kindergarten students' academic knowledge, with an effect size of 0.22). In Reading First schools, the impact of K-PAVE on kindergarten academic knowledge was not statistically significant. This exploratory analysis suggests that the impact of K-PAVE on academic knowledge is different in non-Reading First and Reading First schools (no impact). There was no statistically significant difference in the impact of K-PAVE in Reading First and non-Reading First schools for other outcomes measured at the end of kindergarten or for any outcomes measured at the end of grade 1. For these outcomes, the null hypothesis that the impact of K-PAVE is the same for students in Reading First and non-Reading First schools could not be rejected.

1. INTRODUCTION AND STUDY OVERVIEW

This study is the first randomized study of the impacts of the vocabulary instruction program Kindergarten PAVEd for Success (K-PAVE) (Hamilton and Schwanenflugel 2011) on low-income students in kindergarten and grade 1. The study has two components. The first component is a test of the impacts of one year of the K-PAVE vocabulary instruction in kindergarten, before students have encountered formal reading instruction (the kindergarten study). The second component is a test of whether the effects of the vocabulary instruction in kindergarten are sustained beyond the intervention period, in grade 1 (the grade 1 follow-up study).

The impacts of K-PAVE at the end of the kindergarten intervention year were reported in a previous report on the kindergarten study (Goodson et al. 2010). In that study, K-PAVE was implemented as a supplement to the regular kindergarten curriculum in treatment schools. The primary research question addressed the impact of K-PAVE on students' expressive vocabulary. Secondary research questions addressed impacts on kindergarten students' academic knowledge and listening comprehension. K-PAVE had statistically significant impacts on expressive vocabulary and academic knowledge at the end of kindergarten, with an effect size of 0.14 for each outcome. It did not have a significant effect on listening comprehension. The study also found that K-PAVE had a positive and statistically significant impact on one of three classroom instruction outcomes—vocabulary and comprehension support.

This report presents the findings from the grade 1 follow-up study, which followed students who were part of the kindergarten sample into grade 1. During grade 1, no students received K-PAVE; the study was designed to determine whether the impacts found at the end of kindergarten were sustained in grade 1. The report also presents supplemental findings on impacts at the end of kindergarten and discusses outcomes that were not addressed in the kindergarten report.

ROLE OF VOCABULARY KNOWLEDGE IN READING COMPREHENSION

Identifying effective strategies that schools can use to promote vocabulary acquisition in young, at-risk students is a critical challenge, as many low-income children enter school with limited vocabulary, which has consequences for their subsequent literacy development, especially reading comprehension (Biemiller and Slonim 2001; Coyne, Simmons, Kame'enui, and Stoolmiller 2004). 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 (what is sometimes called *reading vocabulary*). 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, and Adolf 2005; Storch and Whitehurst 2002; Vellutino, Tunmer, Jaccard, and Chen 2007). Many children who 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, and Baldwin 1990; Chall and Conard 1991; Storch and Whitehurst 2002; Spira, Bracken, and Fischel 2005). 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, and 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 (McKeown 1985 [cited in Strickland, Ganske, and Monroe 2002]) and benefit from having new words and concepts that are critical to their learning taught directly to them. (Strickland, Ganske, and Monroe 2002, pp. 108–109)

THE K-PAVE VOCABULARY INSTRUCTION PROGRAM

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, Hamilton, Neuharth-Pritchett, Restrepo, Bradley, and Webb 2010), 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 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. 2010); the components that focus on the alphabet, phonological awareness, and uses of print were not included. Schwanenflugel and her colleagues at the University of Georgia note that, unlike preschool teachers, nearly all kindergarten teachers emphasize these components as part of general literacy instruction.[2] Other alterations to make PAVE appropriate for kindergarten include adaptations to the target vocabulary words and associated books and to teacher training and support activities, to make it practical to take K-PAVE to scale.

Each week of the K-PAVE curriculum is organized around a vocabulary unit consisting of 10 thematically linked target words (see appendix A for a list of K-PAVE materials provided to teachers and appendix B for a sample weekly unit from the K-PAVE program). The target words were selected to align with the themes in the Mississippi state science and social studies frameworks for kindergarten (see appendix C for a list of the 240 K-PAVE target words). 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. For 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.

WHY K-PAVE?

K-PAVE was selected for testing for three primary reasons. First, only a small number of vocabulary interventions are appropriate for this age group. Second, K-PAVE was the only curriculum that had developed teacher training materials and a training protocol. Third, among the vocabulary interventions appropriate for this age group, PAVE (the preschool version of the K-PAVE intervention) was the only one for which an impact study had been completed that provided some evidence of effects. K-PAVE was selected for testing even though it represents an untested modification of the tested curriculum (PAVE) and evidence of its effectiveness was based on a quasi-experimental evaluation with some design limitations (described below).

The study of the original PAVE preschool program evaluated its impact 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. 2010). 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 included 180 boys and 165 girls (n = 350) attending a full-day state prekindergarten program for 4-year-olds (mean age = 4 years, 6 months; standard deviation = 4 months). According to parental report, 56% of participating students were African American, 40% European American, 2% multiracial, 1% Hispanic, 1% Asian American, and 1% not reported. Sixty percent of the children were eligible for free or reduced-price meals. Only children who were native English speakers according to parental

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

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.[3] 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 selection of K-PAVE was based on the results from the preschool study showing an impact of the program on the vocabulary knowledge of young students from low-income households. The results of the preschool study have to be interpreted cautiously, because of design flaws, including potential selection bias in the assignment of schools to condition; small sample sizes; and the inclusion of only a single school in the control group. Although evidence of effects of PAVE was the most important criterion for selecting the program for the current study, it was also important that the curriculum have standardized written instructional materials that could support replication. Compared with the previous study of PAVE, this study provided a stronger test of the curriculum approach and tested that approach when implemented in kindergarten classrooms.

THEORY OF CHANGE FOR K-PAVE

Although improvement in students' vocabulary knowledge at the end of kindergarten is the primary outcome for study, the intervention model extended beyond Explicit Vocabulary Instruction. K-PAVE's focus on conversation and Interactive Book Reading was 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 comprehension of spoken language. Also, as students learn to read, their vocabulary and background knowledge support comprehension of print. For these reasons, K-PAVE was also hypothesized to have impacts on the secondary student outcomes of academic knowledge and listening comprehension.

K-PAVE was also hypothesized to lead to sustained advantages in vocabulary knowledge and other vocabulary-related outcomes through grade 1. The logic model for K-PAVE

[3] 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 race/ethnicity and economic status (as determined by eligibility for free and reduced-price meals) 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.

hypothesizes that increases in vocabulary acquisition and other vocabulary-related outcomes in kindergarten continue to positively affect vocabulary acquisition and other vocabulary-related outcomes in the next school year, even in the absence of the intervention. The logic model hypothesizes that positive effects on vocabulary knowledge at the end of kindergarten would, in grade 1, positively affect the acquisition of early reading skills, specifically passage comprehension.

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

EXPERIMENTAL AND QUASI-EXPERIMENTAL EVIDENCE OF IMPACTS OF VOCABULARY INSTRUCTION

K-PAVE builds on the body of literature on effective strategies for promoting vocabulary knowledge in young students. Although research on the impacts of explicit vocabulary instruction has focused on students in grades 3 and higher (Baumann, Kame'enui, and Ash 2003), gaps in vocabulary evident at school entry (Biemiller and Slonim 2001) underscore the importance of vocabulary instruction before grade 3. The discussion that follows summarizes the evidence from the small (but increasing) set of experimental and quasi-experimental studies in this area that have examined the effectiveness of different approaches to vocabulary instruction.

Two types of vocabulary intervention approaches for preschool and kindergarten children have dominated the research: direct training on word meanings and interactive book reading. Three recent meta-analyses have summarized the research on interventions using *direct vocabulary interventions* with children in preschool and kindergarten to promote vocabulary learning, and the meta-analyses suggest that early elementary school students' vocabulary can be improved through interventions involving these instructional strategies. The National Early Literacy Panel (2008) reported a positive average effect size of .32 for preschool children and .13 for kindergarten children on standardized measures of oral language. Elleman, Lindo, Morphy, and Compton (2009) reported a positive overall effect size of 0.50 of direct vocabulary interventions on comprehension using author-created measures of words taught and a 0.10 effect size for standardized oral language measures. Marulis and Neuman (2010) examined the effects of direct vocabulary interventions on children's word learning. They reported an overall effect size in preschool of 0.85 and in kindergarten of 0.94. They further reported that vocabulary intervention provided by an experimenter was more effective (average effect size of 0.96) compared with vocabulary intervention provided by parents (average effect size of 0.76) or by teachers and child care providers (average effect size of 0.13). Additional findings were that direct vocabulary interventions that combined explicit instruction and implicit instruction (defined as teaching words within the context of an activity, such as storybook reading without intentional stopping) was more effective than either strategy alone (average effect sizes were 1.21 for the combined strategies, 1.11 for explicit instruction alone, and 0.62 for implicit instruction alone). Finally, vocabulary instruction was less effective for low SES children compared with middle to high SES children (average effect sizes were 0.75 and 0.99, respectively).

Another set of meta-analyses that examined the research on *dialogic or interactive book reading*, with teachers, parents and researchers as the participating adult, suggest that embedding vocabulary instruction in repeated reading of the same book can be an effective instructional strategy. The National Early Literacy Panel (2008) reported average effect sizes of 0.66 (kindergarten) and 0.75 (preschool) on oral language outcomes. Mol, Bus, deJong, and Smeets (2008) and Mol, Bus, and deJong (2009) examined studies of interactive book reading for effects on oral language outcomes for children in preschool through first grade. The average effects of dialogic parent-child reading was reported to be 0.50 for preschool and 0.14 for kindergarten. Interactive book reading by an experimenter or teacher were reported to have an average effect size of 0.28.

Individual researchers have examined the effectiveness of different approaches to vocabulary instruction through variants of storybook reading. Exposure to books is a major source of vocabulary learning for children, because books typically contain a wider vocabulary than occurs in conversations (Sulzby, 1985). One set of quasi-experimental studies tested the effect of repeated reading of a story without explicit explanations of word meanings. These studies reported student gains of 5%–9% in the number of instructed words learned at the end of the intervention (Elley 1989; Robbins and Ehri 1994; Hargrave and Senechal 2000; Penno, Wilkinson, and Moore 2002; Brabham and Lynch-Brown 2002; Biemiller and Boote 2006).

A second set of pre-post studies tested the effect of augmenting 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 report an average gain of 12% in instructed words learned (Senechal 1997; Senechal and Cornell 1993). Studies that examined the impact of repeated readings of a story with explanations of word meaning report an average gain of 17% in word meanings known (Robbins and Ehri 1994; Senechal 1997; Senechal, Thomas, and Monker 1995; Hargrave and Senechal 2000; Penno, Wilkinson, and Moore 2002; Brabham and Lynch-Brown 2002; Biemiller and Boote 2006). Some of these studies compare immediate posttests with posttests six weeks to three months after the end of the vocabulary instruction; they report word knowledge to be about the same or higher at delayed testing (Senechal and Cornell 1993; Senechal et al. 1995). 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 and Lynch-Brown 2002; Hargrave and Senechal 2000).

These studies of student gains in vocabulary knowledge of words in storybooks typically used pre-post designs and assessed only gains in student knowledge of the instructed words they were exposed to in the reading texts. Few studies also assessed the effect of vocabulary instruction on untaught vocabulary, by using a standardized receptive or expressive vocabulary measure, for example. Moreover, the sample sizes in the studies tended to be small. These design features limit the strength of the evidence on the impacts of vocabulary instruction on these outcomes for young students.

More recent experimental studies address a weakness of these studies. Repeated reading of the same story by itself does not expose students to additional contexts for increasing their understanding of target words (Beck and McKeown 2007). The experimental studies tested the impact of adding follow-up extension activities or vocabulary reviews to storybook reading 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 statistically significant 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 on 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, every 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 between the storybook and the code-based interventions and 0.85 for the contrast between the storybook intervention and the control group.[4]

Another group of studies tested the effectiveness of instructional approaches that provided opportunities for students to actively participate in book reading, such as through activities in which students engaged 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. Extended vocabulary instruction, or rich instruction, was hypothesized to help students develop greater depth of vocabulary knowledge through exposure to multiple examples of vocabulary words in multiple contexts. This, in turn, was 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.

[4] 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 on experimenter-developed measures, often without a comparison group, and to the fact that research has shown that on average quasi-experiments produce larger estimates of effect sizes than do experiments (Glazerman, Levy, and Meyers 2003).

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 provides 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 study was conducted in eight classrooms in an urban school district with a low-income, predominantly African American population. One classroom from each grade was designated as intervention (implementing Text Talk) and one classroom from each grade was designated 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 (5.6 words) than students in the control group classes (1.0 word) ($p < .001$, 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 assigned to implement the same intervention and half assigned to implement the intervention with additional instructional time. Observations showed that Text Talk alone resulted in 5 encounters per vocabulary word, whereas 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 for students in the enhanced intervention classrooms and 3.8 words for students in the regular intervention classrooms.

Coyne and colleagues 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 on words that received extended instruction than they did on words that received incidental exposure ($p < .001$ for the three outcomes, 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 than they did on words that received embedded instruction (*p* < .001 on all three outcomes, effect sizes of 1.70 for expressive definitions, 0.99 for receptive definitions, and 1.12 for context effect sizes) (Coyne et al. 2007).

A third experimental 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 sizes 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 (effect sizes of 0.97 for receptive definitions and 2.57 for expressive definitions). For embedded instruction, the difference was statistically significant at *p* < .05 for receptive definitions (effect size = 0.24) and at *p* < .01 for expressive definitions (effect size = 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 = 0.38) and partial knowledge (*p* < .01, effect size = 0.56). Mean scores were also 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 for full knowledge and 1.07 for partial knowledge). For embedded instruction, the difference was significant for both full and partial knowledge (effect sizes of 0.63 for full knowledge and 0.49 for partial knowledge).

To date, most of the research on vocabulary instructional strategies has involved small-scale efficacy studies that test 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. As a group, the vocabulary intervention studies do not examine long-term impacts of vocabulary learning. Although there is no empirical basis on which to base a hypothesis about sustained impacts from kindergarten into the next school year, the logic model for K-PAVE posits that increased vocabulary supports subsequent comprehension of reading material and allows students to acquire additional vocabulary through the construction of semantic networks based on the additional vocabulary newly acquired in kindergarten. This follow-up study of K-PAVE examined empirically whether a supplemental vocabulary intervention in kindergarten continued to provide an advantage to students in the next school year, as they began regular reading instruction.

THE FOCUS ON THE MISSISSIPPI DELTA REGION

Research supports the notion 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, and 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 and children from households living in poverty have vocabularies half that size (Hart and Risley 1995). The initial disparity continues through elementary school. By grade 1, children from middle-class households know approximately

5,000 words, and children from households living in poverty know only about 3,000. By grade 4, middle-class children have vocabularies of about 16,000 words, and poor children know only about 11,000 words (White, Graves, and Slater 1990; Beck, McKeown, and Kucan 2002). These trends highlight the need for instructional interventions to accelerate vocabulary acquisition in young children (Biemiller 2001; Catts, Hogan, and Adolf 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 because of the high levels of poverty in the region, students in the Delta have a history of low achievement scores, and the state legislature placed a high priority on meeting the early education needs of students in the Delta (see appendix D for a map of Mississippi showing the Delta region).

The Delta area is primarily rural, with a poverty rate more than twice the national rate (Bishaw and Iceland 2003). According to the U.S. Census Bureau (2008), in 2007 the poverty rate for children under age 18 in the Delta region counties averaged 45.3%, with poverty rates in most counties above 36.0% 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 United States.

For the Southeast Region as a whole, data from several psychometric studies show children from households living in poverty have particularly low vocabulary scores, at about one standard deviation below the national average (Campbell, Bell, and Keith 2001; Restrepo, Schwanenflugel, Blake, Neuharth-Pritchett, Cramer, and Ruston 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 students in grades 3 and 4 in Alabama, Florida, Georgia, Mississippi, and South Carolina do not meet state reading standards. By middle school the proportion is 32%, with higher rates among 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 meeting the early education needs of students through grade 3.

PRIOR FINDINGS AND RESEARCH QUESTIONS

The kindergarten study of K-PAVE is the first large-scale effectiveness trial to test a vocabulary intervention in kindergarten under typical, real-world rather than optimal conditions over the course of a full school year, using a standardized vocabulary outcome measure. The effect of K-PAVE was tested when implemented under real-world conditions in a sample of kindergarten classrooms in multiple school districts in the Mississippi Delta and surrounding areas. Based on a statistical power analysis, we used a kindergarten sample that included 35 school districts, 65 schools, 130 kindergarten classrooms (2 per school), and 1,319 students (about 20 per school). The study used a cluster random assignment design in which schools were assigned either to the K-PAVE intervention as a supplement to the usual classroom instructional program or to the regular instructional program. K-PAVE was offered only during the kindergarten year.

In the follow-up study, the students in the kindergarten sample were followed into grade 1; no students received K-PAVE during grade 1. Although the goals of the K-PAVE intervention are primarily to accelerate kindergarten students' vocabulary development, and secondarily to improve other vocabulary-related outcomes at the end of kindergarten, it was hypothesized that any impacts achieved by the end of kindergarten would carry over into positive effects on students in the next school year, even in the absence of a K-PAVE-style intervention in grade 1. Confirmatory analyses of impacts on students' vocabulary development and other vocabulary-related outcomes at the end of grade 1 were conducted to test this hypothesis. Confirmatory analyses were defined as those that investigated an a priori hypothesis (that is, specified before any examination of the data) with a theoretical or empirical basis. Hypotheses about the sustained impacts of K-PAVE (described further below) were guided by the logic model for the K-PAVE intervention and evidence of short-term impacts of K-PAVE, as reflected in the kindergarten study.

Results of the kindergarten study

In the kindergarten study (Goodson et al. 2010), teachers in the intervention group received training on the K-PAVE intervention in fall of the 2008/09 school year, before the intervention began, in October 2008. Teachers in both the treatment and control groups received their district's usual professional development during the year. Control teachers were offered K-PAVE training the following school year. Kindergarten students in both intervention and control schools were assessed twice during the year—at baseline (fall 2008), before intervention implementation, and again in spring 2009—to examine the impacts of K-PAVE on vocabulary development and related outcomes at the end of kindergarten. Classroom instruction in both intervention and control schools was measured twice during the kindergarten year. Baseline observations were conducted in fall 2008, before K-PAVE training and implementation, and again in spring 2009, to examine the impacts of K-PAVE on kindergarten instructional practices.

The primary research question the kindergarten study addressed was the impact of K-PAVE on kindergarten students' expressive vocabulary. Additional research questions addressed the impacts of K-PAVE on other reading-related outcomes, including 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. It also 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 results of the kindergarten study are as follows:

- 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, and the impact was statistically significant (the 95% confidence interval around the impact estimate was

0.4–2.8 points). The standardized effect size for this impact was 0.14. Translating this effect size into a difference in age–equivalent scores, at the end of kindergarten, students who received the K-PAVE intervention were one month ahead of students in the control group in vocabulary development.

- The impact of K-PAVE on academic knowledge was statistically significant, with a magnitude of 1.95 points on an Item Response Theory–based scale with a mean of 455 points and a standard deviation of 13.5 points in the control group sample (the 95% confidence interval around the impact estimate was 0.2–3.7 points). The standardized effect size for this impact was 0.14. Translating this effect size into a difference in age–equivalent score, at the end of kindergarten, students who received the K-PAVE intervention were one month ahead of students in the control group on academic knowledge.
- K-PAVE did not have a statistically detectable impact on kindergarten listening comprehension.
- K-PAVE had a positive and statistically significant impact on one of three classroom instruction outcomes—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 (0.83 standard deviation) is equivalent to providing comprehension support 12 more times, asking 3 more higher-order questions, and introducing 3 more vocabulary words during a book reading session and to introducing 3 more vocabulary words per hour during other instructional times.[5]
- K-PAVE did not have a statistically detectable impact on instructional support or emotional support in the classroom.
- K-PAVE did not have a statistically detectable impact on the amount of instructional time spent on literacy in areas other than vocabulary and comprehension.

The kindergarten study also examined the fidelity of implementation of K-PAVE among the treatment teachers. Results on fidelity of kindergarten implementation—which are pertinent to the interpretation of carry-over impacts into 1st grade—indicated that the teacher training and support activities were implemented as planned in kindergarten and that, as expected in an effectiveness trial, there was substantial variation in implementation across K-PAVE classrooms. In the classroom, 68% of teachers implemented at least 8 of the 12 instructional strategies with fidelity; 25% of teachers implemented 5–7 strategies with fidelity, and 7% implemented 1–4 strategies with fidelity. No additional measures of implementation fidelity were collected for the 1st grade follow-up, since the studied intervention ended at the end of kindergarten.

Analysis of sustained impacts on students in grade 1

The grade 1 follow-up study provides evidence on whether the effects of K-PAVE in kindergarten led to a sustained advantage in vocabulary and related literacy skills in grade 1, when formal reading instruction typically begins.

[5] The estimated impact of 0.82 standard deviation is equivalent to one more word during each 20-minute period observed during instructional time other than the book read aloud. An additional word during every 20 minutes of instructional time suggests a total of three more words per hour during other instructional times.

The primary research question for the follow-up study was as follows:

1. Is the impact of K-PAVE on students' expressive vocabulary at the end of kindergarten sustained through the end of grade 1?

The study also addresses two additional secondary research questions:

2. Is the impact of K-PAVE on students' academic knowledge at the end of kindergarten sustained through the end of grade 1?
3. Does access to the K-PAVE intervention in kindergarten affect students' passage comprehension in grade 1?

The confirmatory analyses for the grade 1 follow-up study examined the impacts of kindergarten K-PAVE on student outcomes in grade 1 only for outcomes on which there were significant impacts in kindergarten—namely, expressive vocabulary (the primary outcome) and academic knowledge (a secondary outcome). Because there was no impact on listening comprehension in kindergarten, we did not investigate sustained impacts on this measure in grade 1.

The research question about the impact of K-PAVE on passage comprehension at the end of grade 1 was based on the hypothesis that increased vocabulary knowledge at the end of kindergarten leads to impacts at the end of grade 1 that go beyond understanding the meaning of individual vocabulary words to broader passage comprehension. Impacts on passage comprehension were not examined in kindergarten, based on the fact that children typically have not yet been exposed to formal reading instruction on decoding text by the end of kindergarten. Because formal reading instruction has begun by grade 1, we conducted a confirmatory analysis of the impact of access to K-PAVE in kindergarten on students' passage comprehension in grade 1.

Exploratory analysis of different impacts on subgroups of students and schools

This report also presents the results of exploratory analyses of differences in the impact of K-PAVE on subgroups of students and schools in both kindergarten and grade 1. These analyses were considered exploratory for two reasons. First, the study did not have a priori hypotheses about subgroup differences or an empirical basis to guide the investigation of subgroup differences in impacts. The exploratory analyses are intended as an initial investigation of potential differences in impacts that, if found, would need to be investigated in a future study to confirm their accuracy. Second, the subgroup analyses—necessarily based on a partial sample—did not have the same power as the main impact analysis and therefore may not be able to detect subgroup differences that do exist.

The study addressed two exploratory research questions about subgroup differences in effects:

4. Do the impacts of K-PAVE on students' expressive vocabulary, academic knowledge, and listening comprehension at the end of kindergarten and on students' expressive vocabulary, academic knowledge, and passage comprehension at the end of grade 1 vary by student gender and pretest score?
5. Do the impacts of K-PAVE on students' expressive vocabulary, academic knowledge, and listening comprehension at the end of kindergarten and on students' expressive vocabulary, academic knowledge, and passage comprehension at the end of grade 1 differ in Reading First and non-Reading First schools?

Exploratory analysis of kindergarten impacts on other outcomes

Two other analyses were conducted to examine additional impacts of K-PAVE in kindergarten. These analyses were not included in the main kindergarten report because they are considered exploratory. One set of analyses investigated the impacts of K-PAVE on the four discrete teacher instructional practices that were combined in the broad measure of vocabulary and comprehension support that was tested as an outcome in the kindergarten confirmatory impact analysis. This analysis was considered exploratory because it was undertaken only after a significant impact was found on the combined construct. Analyses of the four component variables were not specified as part of the analysis plan. These analyses were undertaken to help better understand which aspects of vocabulary and comprehension support are improved by K-PAVE. Specifically, they addressed the following research question:

6. What is the impact of K-PAVE at the end of the intervention on each of the four components of vocabulary and comprehension support in the classroom (introduction of new vocabulary in the classroom during book read alouds, introduction of new vocabulary in the classroom during other instructional time, provision of comprehension support during book read alouds, and use of open-ended questions during book read-alouds)?

The second set of exploratory analyses addressed questions about the impact of K-PAVE at the end of kindergarten on students' and teachers' lexical diversity, an alternative measure of vocabulary production. Lexical diversity is a measure of vocabulary use in oral language, derived from the ratio of the number of unique words used in a language sample to the total number of words used. The lexical diversity outcomes were categorized as exploratory analyses in the study plan because the measure itself was new and information on the psychometric properties of the measure for adults or young children is not available. Moreover, the measure of student lexical diversity was measured for only 40% of the student sample, which led us to expect that the analysis would be underpowered.[6] This second set of exploratory analyses addressed the following questions:

7. What is the impact of K-PAVE on lexical diversity in students' elicited language production, measured at the end of kindergarten?
8. What is the impact of K-PAVE on lexical diversity in teachers' naturally occurring language production, measured at the end of the K-PAVE intervention period?

ORGANIZATION OF THE REPORT

The remainder of the report details the study design and methodology of the follow-up study and examines its impact. 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 presents the results of the confirmatory impact analyses of student outcomes in grade 1, one year after the end of the K-PAVE intervention. Chapter 4 discusses the exploratory analyses of subgroup differences in impacts in kindergarten and grade 1. Appendix E presents the results of additional exploratory analyses of the impacts of K-PAVE on the components of the vocabulary and comprehension support composite measured in kindergarten and on student and teacher lexical diversity in kindergarten. Additional appendixes provide more information about measures, statistical models, imputation of missing data, and sensitivity analyses.

[6] To keep costs down, we asked only 40% of the student sample to complete the Elicited Language Task, from which the lexical diversity measure was created. Four students in each class (eight students per school) were administered the task; these students were randomly selected from the sample of study students from that class. Of the 1,296 students in the full sample, 521 (40.2%) were selected to complete the Elicited Language Task (244 students in treatment schools and 277 students in control schools). Data were actually collected from 233 students (95%) in treatment schools and 262 students (95%) in control schools.

2. Study design and methodology

This report is a follow-up study examining the impact of the Kindergarten PAVEd for Success (K-PAVE) intervention on students' vocabulary in grade 1. The study used a cluster random assignment design in which schools were randomized to intervention or control conditions.

A previous report (Goodson et al. 2010) presented findings on the impact of K-PAVE on student outcomes and classroom instruction at the end of the kindergarten intervention year. K-PAVE was found to have positive and statistically significant impacts on kindergarten students' expressive vocabulary (effect size = 0.14) and academic knowledge (effect size = 0.14) and on classroom vocabulary and comprehension support in kindergarten (effect size = 0.83). K-PAVE was not found to have a statistically detectable impact on kindergarten students' listening comprehension, classroom instructional support, or classroom emotional support.

This report examines whether the impacts of K-PAVE were sustained in grade 1, one year after the end of the intervention. It relies on the same rigorous experimental design of the kindergarten impact study, which addresses the limitations of earlier research noted in chapter 1 and provides a valid basis for answering the study's key research questions about sustained impacts of K-PAVE.

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 to April 2009; kindergarten teachers in intervention schools were trained in 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 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.[7] Control teachers were offered the K-PAVE intervention training the following school year.[8] Because of the experimental design of the study, differences between intervention schools and control schools in student outcomes (including those measured one year after the end of the intervention) and classroom instruction that exceed chance (that is, are statistically significant) can be attributed to the K-PAVE intervention.

Assigning all classes in a school to the same condition eliminated concerns about potential diffusion of effects across classrooms within a school, which can occur when classrooms are randomly assigned within each school (through discussion, observation, or other

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

[8] 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. Grade 1 teachers do 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 thus not likely to implement the K-PAVE program in their classrooms.

forms of cross-teacher communication). Implementing the intervention in all kindergarten classrooms in an intervention school was also 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 tested 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 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.[9] 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 with similar characteristics, including high rates of poverty, low student achievement, and predominantly rural and African American communities (see map in appendix D). The target population for the study was all elementary schools in the Mississippi Delta region with at least two kindergarten classes. The Delta Revitalization Act of 2006 identified the counties that comprise the Mississippi Delta region.

Recruitment of eligible districts, schools, and teachers

Informed by the statistical power analysis conducted in the evaluation design phase, we set a recruitment target based on the goal of following students for one year after the intervention. The recruitment target was 60–70 schools, with 2 classrooms per school and 10 students per classroom. (See appendix F for 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):

- 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%).
- The district has at least one school with two kindergarten classes and is in a high-poverty community (with at least 40% of students eligible for free or reduced-price meals).

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

- The district has at least one school that meets these criteria and is willing and able to allow 2 consenting kindergarten teachers to be randomly selected for observation and 20 students to be randomly selected for testing from among students with parental permission.
- 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 2 kindergarten classrooms. To ensure that a sample of at least 60 schools could be obtained (the sample size goal based on power calculations), we expanded the sampling frame to include schools from districts contiguous and sharing demographic characteristics with the Mississippi Delta region.[10] 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 = 97%), on average, than schools in the surrounding region (median = 84%; $t = –2.28$, $p = .03$).

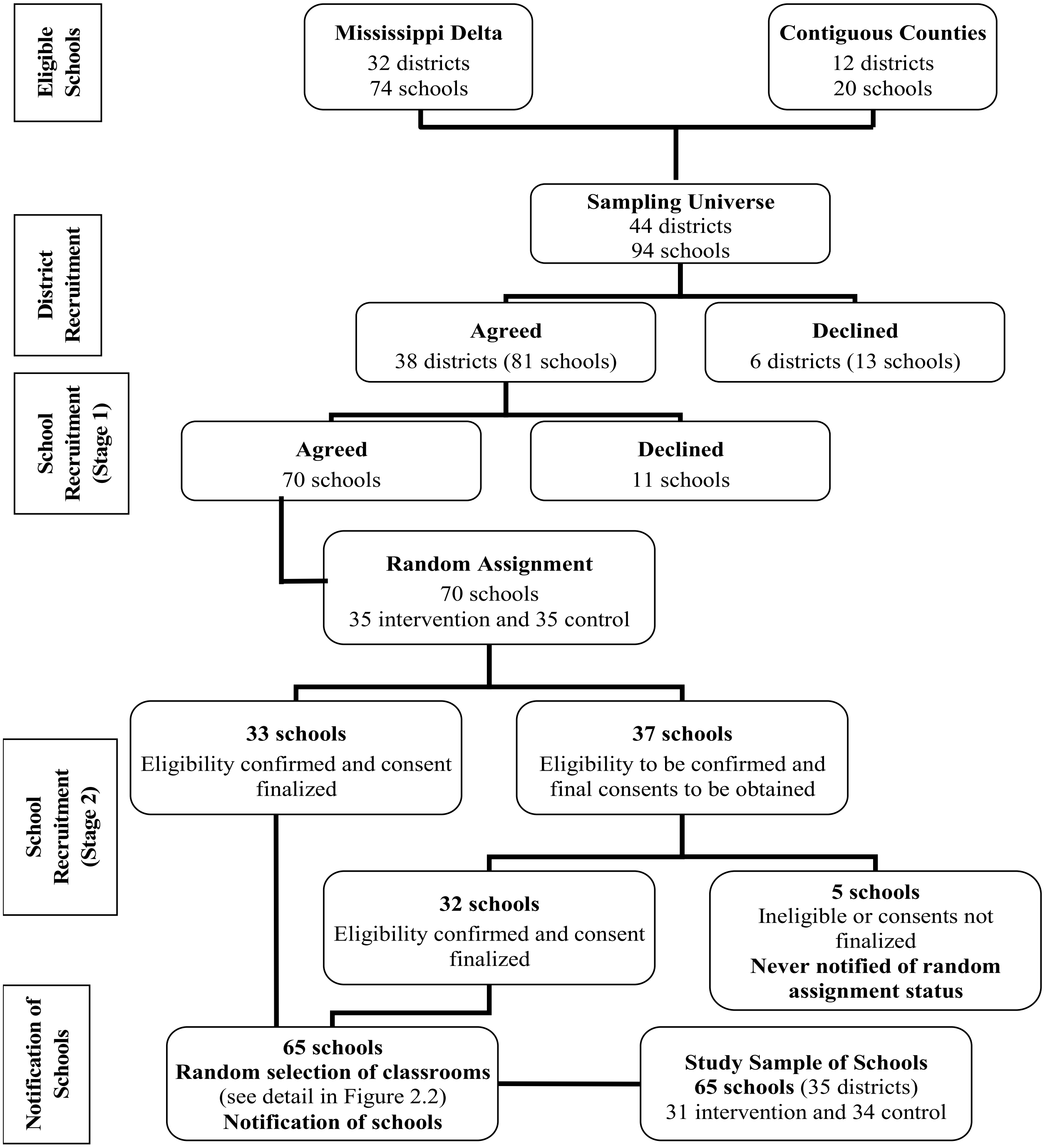
Recruitment of districts, schools, and teachers took place over January–August 2008. Of the 44 district superintendents (86%), 38 agreed to participate. Once a district superintendent agreed to participate, principals of all eligible schools in the district were recruited for the study. Once schools 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 (figure 2.1). Schools were recruited between February and June 2008; random assignment of schools occurred at the end of July. 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 their willingness to participate; obtain consent for all kindergarten

[10] 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 (less 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 midsize 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. It started with a list of districts and elementary schools in contiguous counties in which at least 40% of students were eligible for free or reduced-price meals. Based on its insider's perspective, it determined whether each school was similar to those in the Delta in demographic, poverty, and achievement data. Based on its ratings, 13 additional schools were eliminated from the potential sampling frame.

teachers, including any teachers hired since the first phase of recruitment; and obtain any outstanding written consent forms not yet submitted by the school.

Figure 2.1. District and school recruitment process and timeline


Eligible Schools
Mississippi Delta
32 districts
74 schools
Contiguous Counties
12 districts
20 schools
District Recruitment
Sampling Universe
44 districts
94 schools
Agreed
38 districts (81 schools)
Declined
6 districts (13 schools)
School Recruitment (Stage 1)
Agreed
70 schools
Declined
11 schools
Random Assignment
70 schools
35 intervention and 35 control
School Recruitment (Stage 2)
33 schools
Eligibility confirmed and consent finalized
37 schools
Eligibility to be confirmed and final consents to be obtained
32 schools
Eligibility confirmed and consent finalized
5 schools
Ineligible or consents not finalized
Never notified of random assignment status
Notification of Schools
65 schools
Random selection of classrooms
(see detail in Figure 2.2)
Notification of schools
Study Sample of Schools
65 schools (35 districts)
31 intervention and 34 control

Before randomization, 70 of the 81 eligible schools indicated their willingness to participate in the study. All 70 schools were included in the pool of schools to be randomized, regardless of whether they had submitted all of the written consent forms. Schools were randomized on July 24, 2008, with 35 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; on July 25, 2008, letters were sent notifying these 33 schools of their random assignment status and identifying 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, all staffing was confirmed, and all of the necessary written consent forms had been submitted for 32 of the remaining 37 schools. The other five schools were excluded from the study because they did not meet eligibility criteria or were unwilling to participate in the study.

Although all of 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, as the schools did not know the group to which they had been assigned).

The final sample included 65 schools, including 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 and 34 control schools.[11]

Random assignment within blocks

Schools were placed into three blocks based on previous participation in reading initiatives.[12] Among the 65 schools in the study sample, 17 were Reading First schools, 5 had a

[11] 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 consent) before notification of random assignment.

[12] Although within-district random assignment would have controlled for district characteristics, doing so would not have resulted in a large enough 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 region, 15 had only one elementary school with kindergarten. A within-district random assignment design would have reduced the sampling frame 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

Mississippi state reading initiative (Barksdale Reading or Mississippi Sufficiency), and 43 had neither Reading First nor a state reading initiative. Schools in the latter group may have had a local district initiative or no reading initiative, but information on the existence of such reading initiatives was not collected as part of this study. Within the reading initiative blocks, schools were matched based on a set of school characteristics: the School Performance Classification;[13] the percentage of students eligible for free or reduced-price meals; the percentage of African American students; the locale type (rural, small town, large town or fringe of city); and the location (Delta or contiguous county). (Appendix G, which details the process of random assignment, includes a description of the matching within blocks.) Once matched based on these school characteristics, schools were randomly assigned to intervention or control conditions.

Random selection of classrooms

The 65 schools in the sample had a total of 256 full-day kindergarten classrooms (figure 2.2). The average consent rate for teachers across the 65 schools was 91.5%.[14] Because at least two teachers from all 65 schools consented to participate in the study, no schools were lost because of 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.[15] This random selection ensured that selected teachers were representative of teachers in their schools who were willing to participate in the study. For schools with only two kindergarten classrooms or only two consenting kindergarten teachers (40% of schools), both

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 heterogeneous within districts. The comparatively small number of schools and potentially large variation within districts would tend to result in less comparable intervention and control groups than random assignment within blocks based on prior experience with reading initiatives.

[13] 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 (that is, minimal, basic, proficient, 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"; "proficient" is defined as "solid academic performance and mastery of the content area knowledge and skills required for success at the next grade" (Mississippi Department of Education, n.d.). A school's achievement level is 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.

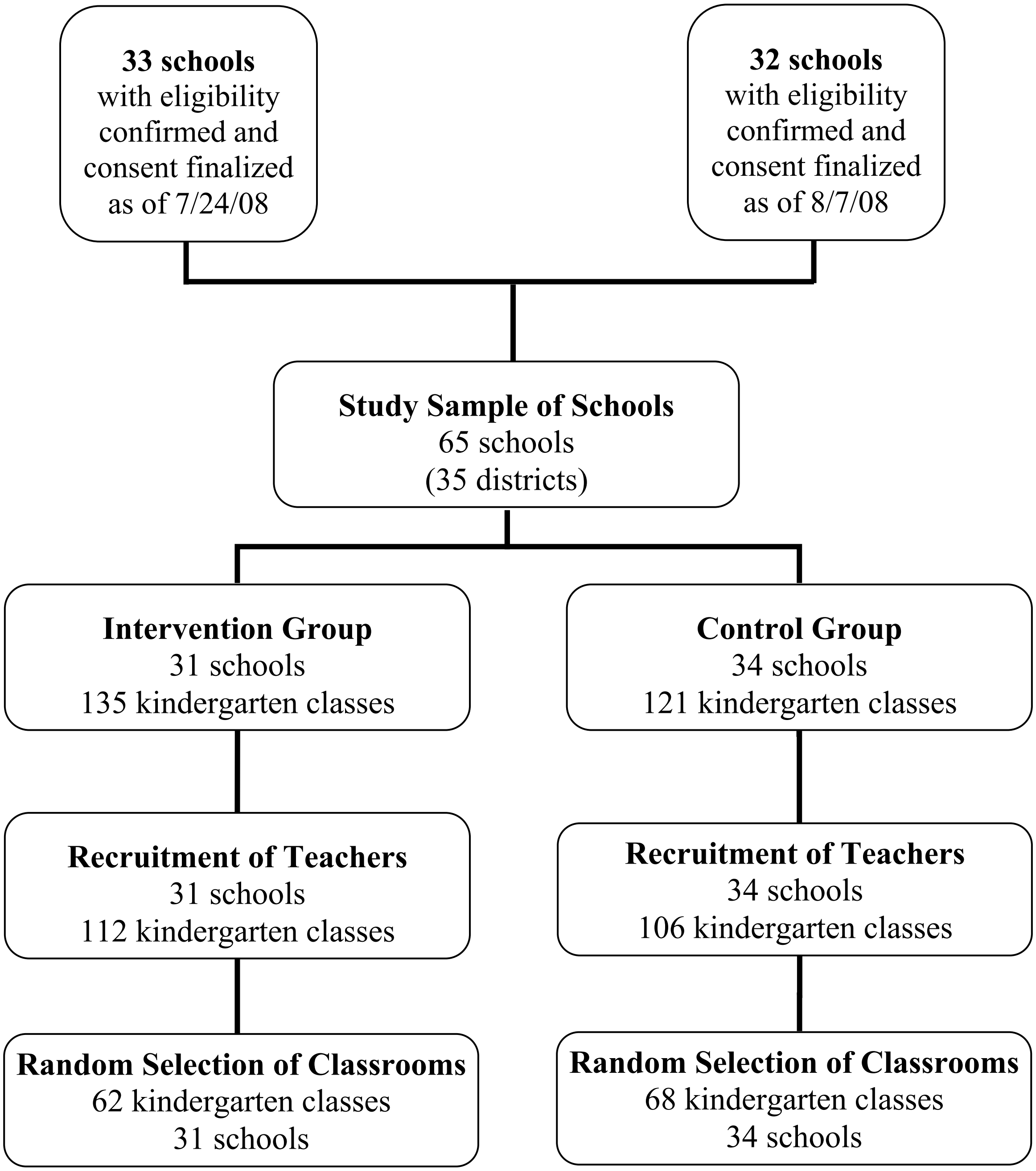
[14] Teacher consent rates ranged from 20% to 100%. In 74% of schools, all kindergarten teachers consented to be in the study.

[15] Sixty-six percent of schools had more than two kindergarten classrooms: 31% of schools had three kindergarten classrooms; 29% had 4–8 classrooms, and 6% had 11–15 classrooms.

classrooms were selected. Schools were not notified of their assignment condition until after the random selection of teachers was completed.

In 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 teachers participated in the professional development activities.

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


33 schools with eligibility confirmed and consent finalized as of 7/24/08
32 schools with eligibility confirmed and consent finalized as of 8/7/08
Study Sample of Schools 65 schools (35 districts)
Intervention Group 31 schools 135 kindergarten classes
Control Group 34 schools 121 kindergarten classes
Recruitment of Teachers 31 schools 112 kindergarten classes
Recruitment of Teachers 34 schools 106 kindergarten classes
Random Selection of Classrooms 62 kindergarten classes 31 schools
Random Selection of Classrooms 68 kindergarten classes 34 schools

Random selection of students

Based on statistical power calculations, we 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, we 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 spring.[16] Theoretically, parents' decision to permit their children to be assessed could have been influenced by the school's intervention status, as recruitment of students occurred after 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 received 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 were full-day classrooms enrolling an average of 20 students each, with enrollment ranging from 12 to 27 students. Parental permission was received for an average of 14 students per classroom (73% average permission rate), with the number of permissions ranging from 3 to 25 students per classroom (12–48 students per school). Only 11 schools (17%) received permission from fewer than 20 students from the two selected classrooms combined. On average, 20 students per school were randomly selected to be assessed (74% average selection rate per school). From among 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 H 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 later in the chapter, which illustrates the flow of students through the study.) At the time of baseline data collection, 46 of the randomly selected students (23 in intervention schools and 23 in control schools) were not in school.[17] In their place, 43 alternates (21 in intervention schools and 22 in

[16] Although school begins the first week of August in Mississippi, 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.

[17] Appendix I compare 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

control schools) were randomly selected for testing. The final analytic sample includes all of the 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 final study sample

Comparing study schools with those that declined to participate. The sample of schools that volunteered for the study was similar to the schools that declined to participate, in terms of School Performance Classification, expectations for annual growth in student achievement, and the percentage of students eligible for free or reduced-price meals (table 2.1). The sample schools differed from those that declined to participate on two characteristics: sample schools had a greater percentage of African American students (mean = 85%) than schools that declined to participate (mean = 73%, $t = –2.17$, $p = .03$), and a larger percentage of sample schools than schools that declined were located in small towns and a smaller percentage in rural areas (x^2 = 6.53, $p = .03$).

Table 2.1. Characteristics of schools that agreed to participate in study, schools that declined to participate in the study, and all eligible schools

(percent, except where otherwise indicated)

Characteristic	Agreed (study sample) (*n* = 65 schools)	Declined (*n* = 28 schools)	Sampling frame (*n* = 93 schools)	Test of difference[a]
Location				$\chi^2 = 2.69$, $p = .10$
Within the Delta	83.1	67.9	78.5	
Contiguous to the Delta	16.9	32.1	21.5	
School Performance Classification (grade 3 and higher)				$\chi^2 = 4.30$, $p = .12$
Low or underperforming	20.7	30.8	23.8	
Successful	55.2	30.8	47.6	
Exemplary or superior	24.1	38.5	28.6	
Annual growth expectation (from previous school year for grades 3 and higher				$p = .75$
Met or exceeded	17.2	11.5	15.5	
Not met	82.8	88.5	84.5	
Percentage of students eligible 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 percent	16.7	32.1	21.5	
Mean (standard deviation)	85.2 (16.3)	78.5 (19.0)	83.2 (17.3)	$t = –1.71$, $p = .09$

for free or reduced-price meals, Individualized Education Programs status, or 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.

Characteristic	Agreed (study sample) (*n* = 65 schools)	Declined (*n* = 28 schools)	Sampling frame (*n* = 93 schools)	Test of difference[a]
Percentage of African American students				
96–100	53.9	32.1	47.3	
81–95	21.5	17.9	20.4	
Less than 81 percent	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 type				χ^2 = 6.53, *p* = .03
Rural	47.7	67.9	53.8	
Small town	36.9	10.7	29.0	
Large town or fringe of midsize city	15.4	21.4	17.2	

Note: Distributions of school characteristics for cases with missing data were assumed to be the same as for cases with nonmissing data. Rates of missing data ranged from 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 the percentage of African American students; although categories are presented for these variables in the table, they are continuous variables. Fisher's exact test was used to test for differences in annual growth expectation because of small cell sizes. For Fisher's exact test, only *p*-values and no test statistics are reported. For chi-square tests and *t*-tests, both *p*-values and test statistics are reported.

Description of study schools. 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 neither Reading First nor a state reading initiative (see table 2.2). Schools in the latter group may have had a local reading initiative or no initiative, but information on the existence of such reading initiatives was not collected as part of this study. On the School Performance Classification for 2006/07, 21% of schools were classified as low-performing or underperforming, 55% were classified as successful, and 24% 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); the median percentage of students eligible for free or reduced-price meals was 92% (not shown). Forty-eight percent of the schools were in rural areas, 37% in small towns, and 15% in large towns or on the fringe of a city.

Comparison of intervention and control schools. 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

(percent, except where otherwise indicated)

Characteristic	Control group (*n* = 34 schools)	Intervention group (*n* = 31 schools)	Full sample (*n* = 65 schools)	Test of difference[a]
Reading initiatives				*p* = .99
Reading First/state reading initiative	35.3	32.3	33.9	
Local or no initiatives	64.7	67.7	66.2	
School Performance Classification (grades 3 and higher				*p* = .92
Low or underperforming	19.4	22.2	20.7	
Successful	58.1	51.9	55.2	
Exemplary or superior	22.6	25.9	24.1	
Annual growth expectation (from the previous school year for grades 3 and higher)				*p* = .99
Met or exceeded	16.1	18.5	17.2	
Not met	83.9	81.5	82.8	
Percentage of students eligible 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 percent	17.7	16.1	16.7	
Mean (standard deviation)	84.8 (15.5)	85.6 (17.4)	85.2 (16.3)	*t* = 0.96, *p* = .34
Percentage of African American students				
96–100	52.9	54.8	53.9	
81–95	20.6	22.6	21.5	
Less than 81 percent	26.5	22.6	24.6	
Mean (standard deviation)	84.0 (15.2)	86.0 (20.3)	85.0 (22.8)	*t* = –0.34, *p* = .74
Locale type				χ^2 = 0.96, *p* = .62
Rural	47.0	48.4	47.7	
Small town	41.1	32.3	36.9	
Large town or fringe of midsize city	11.8	19.4	15.4	
Location				χ^2 = 0.68, *p* = .41
Within the Delta	79.4	87.1	83.1	
Contiguous to the Delta	20.6	12.9	16.9	

Note: Distributions of school characteristics were assumed to be the same for cases with missing data as for cases with non-missing data. Rates of missing data ranged from 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; for Fischer's exact tests, only *p*-values and no test statistics are reported. *t*-tests were used to test for intervention and control group differences in eligibility for free or reduced-price meals and the percentage of African American students; although categories are presented in the table for these variables, they are continuous variables. For *t*-tests, both *p*-values and test statistics are reported. Chi-square tests were used to test for intervention and control group differences in locale type and location; both *p*-values and test statistics are reported.

Schools were blocked by reading initiative to ensure that intervention and control schools were balanced on this factor. This step was taken because we expected that experience with Reading First or a state reading initiative would be related to classroom instructional practices, student outcomes, or both. By ensuring that intervention and control schools were balanced with regard to reading initiatives, we were able to conclude that any impacts detected could be attributed to K-PAVE rather than to the Reading First program or a state reading program. Because experience with other reading initiatives may interact with how teachers implement K-PAVE and with its effectiveness for improving students' vocabulary outcomes, we conducted an exploratory subgroup analysis to examine whether impacts of K-PAVE differed depending on whether schools used Reading First (see chapter 4).

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 (χ^2 = 2.51, p = .47) (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

(percent, except where otherwise indicated)

Publisher/reading series	**Intervention schools (n = 30)**	**Control schools (n = 34)**	**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 (includes nine other curricula)	30.7	35.3	

Source: Telephone survey of schools.

ATTRITION AND ANALYTIC SAMPLE IN THE KINDERGARTEN STUDY

Attrition of schools and classrooms

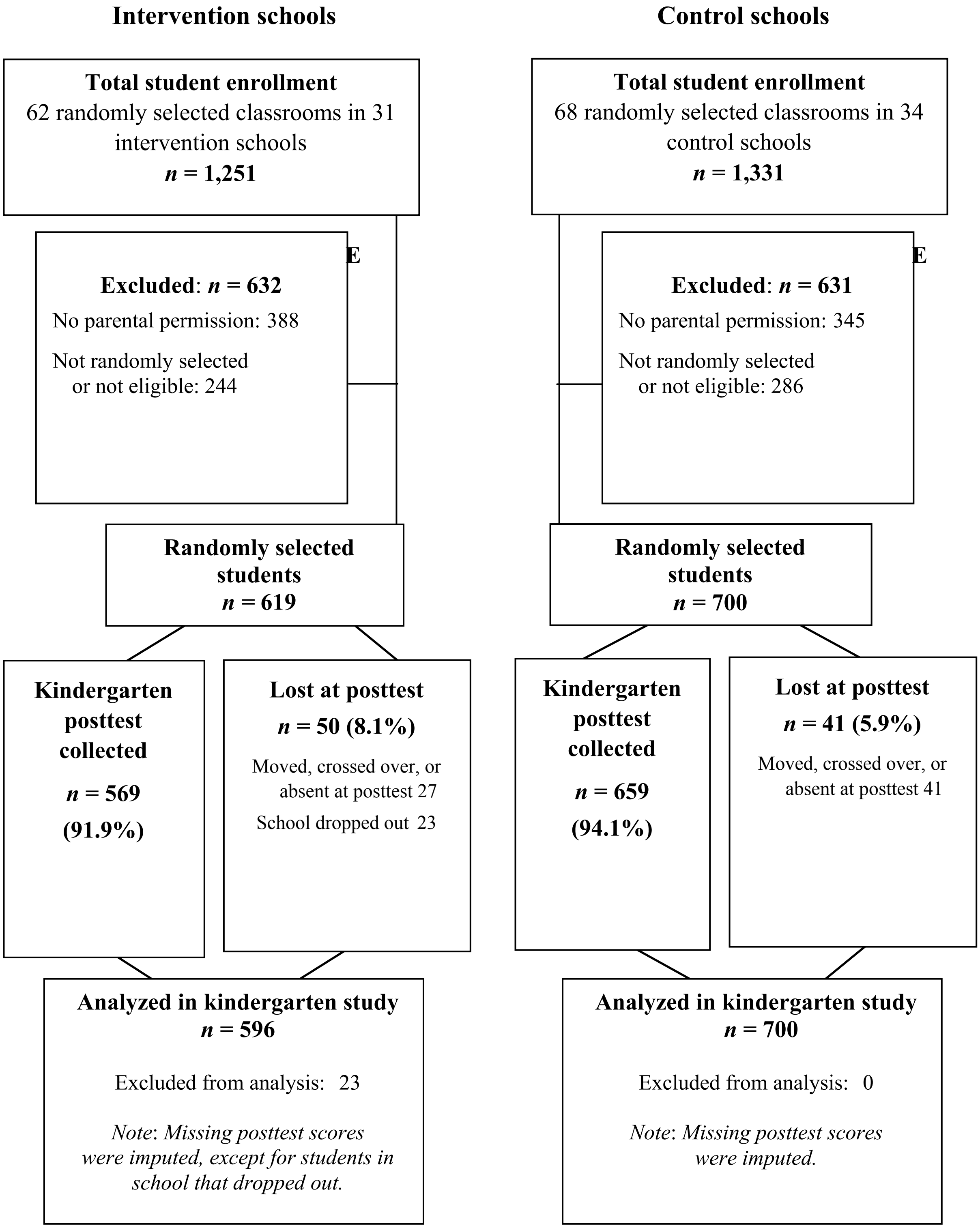
At the time of random assignment, the sample included 65 schools, 130 classrooms, and 1,319 students. One treatment school dropped out of the study during the intervention year and did not provide data for impact analysis, resulting in an overall school-level attrition rate of 1.5% and a differential attrition rate of 3.0% (the school attrition rate was 3.0% for intervention schools and 0 for control schools). This school did not differ from the average school in the intervention group or the average school in the sample. 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.

The school that was lost was in the block of intervention schools that had neither the Reading First program nor a Mississippi reading initiative. The remaining schools in the block were weighted to adjust for the loss of the school from this block (see appendix J for a discussion of the weighting used), and models were estimated without weights to examine sensitivity to weighting (see appendix K).

Attrition of students

The flow of students through the kindergarten study—including recruitment, random selection, and attrition from data collection—is shown in figure 2.3. At the kindergarten 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 was 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 were included ($\chi^2 = 2.19$, $p = .14$). With the exception of the one intervention school that dropped out of the study, students left the study because they moved out of state, transferred to a nonstudy school in Mississippi, or were absent during the data collectors' visits. (See the section later in this chapter on data analysis methods and appendix L for details on the imputation of missing data. Appendix K reports sensitivity analyses examining the influence of missing data imputation on impact estimates and their standard errors.)

Figure 2.3. Flow of students through the kindergarten study


Intervention schools
Control schools
Total student enrollment
62 randomly selected classrooms in 31 intervention schools
n = 1,251
Total student enrollment
68 randomly selected classrooms in 34 control schools
n = 1,331
E
Excluded: *n* = 632
No parental permission: 388
Not randomly selected or not eligible: 244
E
Excluded: *n* = 631
No parental permission: 345
Not randomly selected or not eligible: 286
Randomly selected students
n = 619
Randomly selected students
n = 700
Kindergarten posttest collected
n = 569
(91.9%)
Lost at posttest
n = 50 (8.1%)
Moved, crossed over, or absent at posttest 27
School dropped out 23
Kindergarten posttest collected
n = 659
(94.1%)
Lost at posttest
n = 41 (5.9%)
Moved, crossed over, or absent at posttest 41
Analyzed in kindergarten study
n = 596
Excluded from analysis: 23
Note: *Missing posttest scores were imputed, except for students in school that dropped out.*
Analyzed in kindergarten study
n = 700
Excluded from analysis: 0
Note: *Missing posttest scores were imputed.*

Note: Student crossovers are listed as "lost at posttest" because they do not receive the full duration of their assigned condition; however, posttest data for student crossovers were collected and analyzed.

Nonparticipants and crossovers during the kindergarten intervention year

Students. Forty-five students were categorized as nonparticipants in data collection at the end of kindergarten—students who moved out of state or transferred to a nonstudy school in Mississippi during the intervention year—either before baseline testing or between pretest and posttest. Of the 45 students, 21 were in intervention schools and 24 in control schools; there was no statistically significant difference in the rate of nonparticipation by condition ($\chi^2 = 0.009$, $p = .93$). To maintain the statistical equivalence of the two groups, all students (except students in the school that dropped out of the study) were included in the analysis. However, nonparticipants had no posttest data. To retain nonparticipants in the analysis, values were imputed for missing pretest and posttest data (see appendix L).

Nonparticipants in the intervention group did not experience the full length of the intervention. Therefore, their inclusion in the analysis could diminish the estimated effect of the K-PAVE intervention—particularly to the extent that some of these students would have received a full year of K-PAVE in a statewide program. However, sensitivity analyses conducted as part of the kindergarten study indicated that impacts of K-PAVE on kindergarten students' outcomes and classroom instruction at the end of the intervention year were not affected by whether nonparticipants were kept in the sample (see appendix K for sensitivity analyses).

There were only five crossover students—students initially enrolled in a control school who switched to an intervention school during the intervention year or students initially enrolled in an intervention school who switched to a control school during the intervention year. The crossover rate was not significantly different for the intervention and control schools ($p = .99$).[18] Many crossovers can compromise the integrity of the impact estimates. The low rate of crossovers in this study did not appear to threaten the validity of the impact estimates in the kindergarten study. To preserve the integrity of the random assignment, crossover students were analyzed in their original assigned condition. Sensitivity analyses were conducted as part of the kindergarten study. The impacts of K-PAVE on kindergarten students' outcomes and classroom instruction at the end of the intervention year were not affected by the inclusion or exclusion of the crossover students (see appendix K).

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.

ATTRITION AND ANALYTIC SAMPLE IN THE GRADE 1 FOLLOW-UP STUDY

For the grade 1 follow-up study, the study design included assessment of all students who were in the same school in school year 2009/10 as they were at the time of random selection in kindergarten (school year 2008/09) or who had transferred to a different school that was in the original study sample. Students were assessed at follow-up regardless of their grade level; students retained in kindergarten as well as those who had moved to grade 1 in school year

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

2009/10 were tested at follow-up.[19] We did not follow students who moved out of state, transferred to nonstudy schools in Mississippi, or transferred out of their kindergarten school to a school that was unknown to the kindergarten school staff. Although it is possible that students who moved out of state or to nonstudy schools came from families that were qualitatively different from other families in the study (possibly more residentially unstable or economically more upwardly mobile), the rate of attrition was similar for intervention and control students. As the likelihood of family relocation was not likely to have been affected by the vocabulary curriculum in the child's school, excluding relocated students from the sample should not create treatment-control group differences that bias the estimated impact of K-PAVE at the end of grade 1.

As shown in table 2.4, 1,132 students (87.3%) in the kindergarten sample were assessed at follow-up in grade 1, including 69% of the students in the kindergarten analytic sample who moved to grade 1 but remained in the same school, 14% of students in the kindergarten sample who moved to grade 1 in other schools in the study sample, and 4% of students in the kindergarten analytic sample who were retained in kindergarten in a study school. The remaining 13% of the kindergarten sample moved out of state, moved to nonstudy schools in the state, or were absent at follow-up. Student attrition was 12.4% in the intervention group and 12.9% in the control group. There was no statistically significant difference in these attrition rates ($\chi^2_{(1)}$ = 0.06, p = .81).

Table 2.4. Student mobility from kindergarten to grade 1

	Control schools		Intervention schools		Total		Test of difference[a]
Student group (*n* = 1,296)	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**	
Analytic sample	700	100.0	596	100.0	1,296	100.0	$\chi^2_{(1)}$ = 0.06 = .81
Tested at follow-up[b]	610	87.1	522	87.6	1,132	87.3	
No transfer	504	72.0	446	74.8	950	73.3	
Transfer to study school	106	15.1	76	12.8	182	14.0	
Not tested at follow-up[c]	90	12.9	74	12.4	164	12.7	

a. The chi-square test is for a 2 x 2 table (study condition x follow-up status).

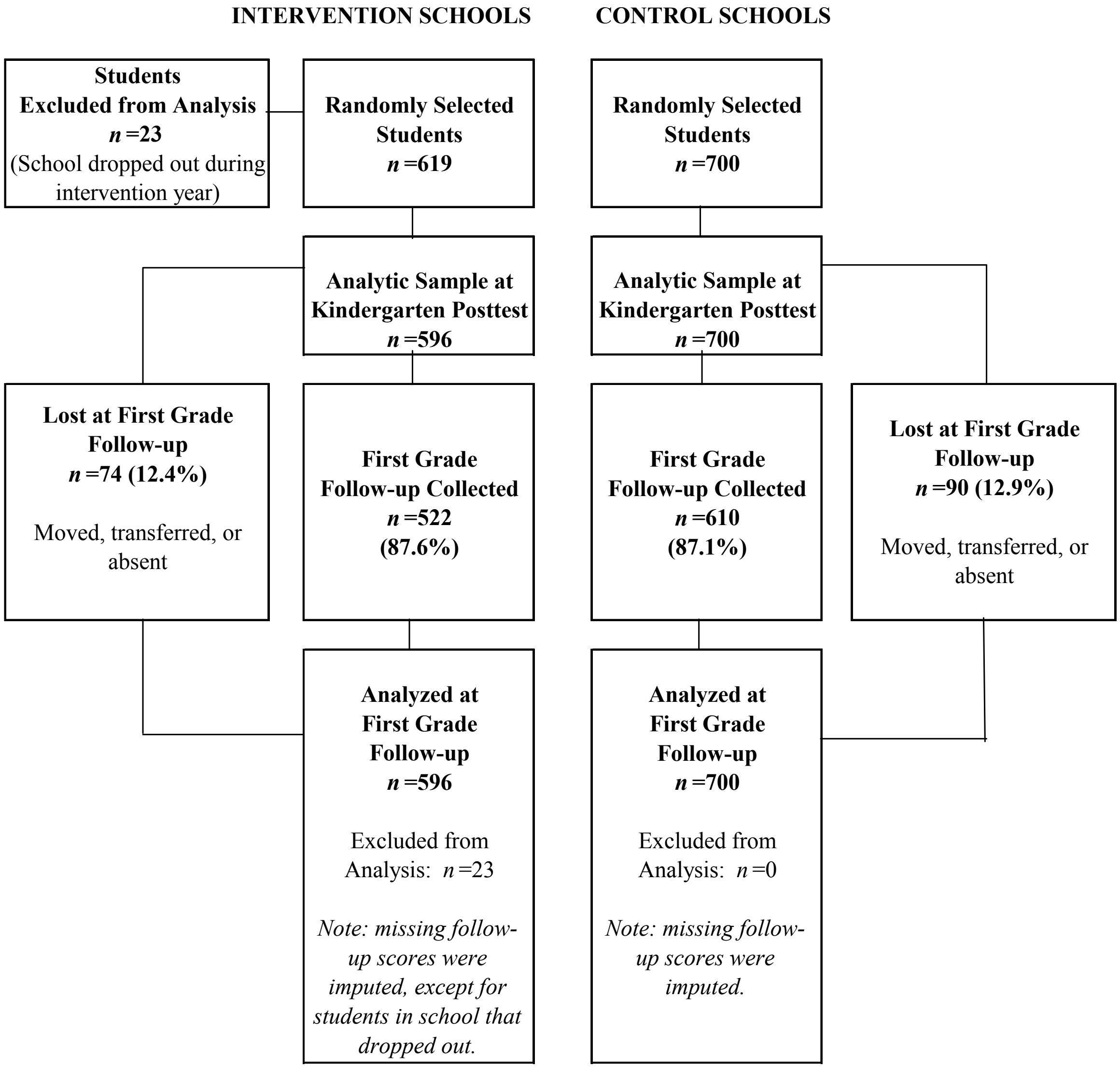
b. Students who remained in their original school or transferred to another study school.

c. Students who transferred to nonstudy schools, transferred out of state, or were absent at follow-up.

[19] Fifty-seven students in the analytic sample (4.4%) were retained in kindergarten in study schools. Students retained in kindergarten in control schools may have received K-PAVE during their second kindergarten year, because control teachers were offered the K-PAVE training after the first year of the study. Students who remain in kindergarten for a second year in an intervention school may have been exposed to two years of K-PAVE. With equal shares retained (4.2% in intervention and 4.9% in control; χ^2 = 0.35, p = .55), the single year of K-PAVE received by the control students retained in kindergarten should roughly offset the second year received by the intervention students retained in kindergarten, causing no overall bias.

Figure 2.4 shows the flow of students through the follow-up year of the study. For both intervention and control groups, it shows the randomly selected sample of students, the analytic sample at kindergarten posttest, the number of students lost at grade 1 follow-up, the number of students on whom grade 1 follow-up data were collected, and the analytic sample at grade 1 follow-up.[20]

Figure 2.4. Flow of students through grade 1 year of the study


INTERVENTION SCHOOLS
CONTROL SCHOOLS
Students Excluded from Analysis n =23 (School dropped out during intervention year)
Randomly Selected Students n =619
Randomly Selected Students n =700
Analytic Sample at Kindergarten Posttest n =596
Analytic Sample at Kindergarten Posttest n =700
Lost at First Grade Follow-up n =74 (12.4%) Moved, transferred, or absent
First Grade Follow-up Collected n =522 (87.6%)
First Grade Follow-up Collected n =610 (87.1%)
Lost at First Grade Follow-up n =90 (12.9%) Moved, transferred, or absent
Analyzed at First Grade Follow-up n =596 Excluded from Analysis: n =23 Note: missing follow-up scores were imputed, except for students in school that dropped out.
Analyzed at First Grade Follow-up n =700 Excluded from Analysis: n =0 Note: missing follow-up scores were imputed.

[20] The analytic sample at grade 1 follow-up included all students in the analytic sample at kindergarten posttest. Missing scores at grade 1 follow-up were imputed. The section on data analysis methods in this chapter and appendix L provide details on the imputation of missing data. Appendix K reports sensitivity analyses examining the influence of missing data imputation on impact estimates and their standard errors.

MEASURES AND DATA COLLECTION

Analysis of the confirmatory and exploratory research questions was based on data on student performance collected at baseline and at the end of grade 1, as well as covariate data on student characteristics and school characteristics. To address the exploratory research questions about subgroup differences in kindergarten impacts, we based the analyses on the same data collected at baseline and on data on student performance at the end of kindergarten. Table 2.5 shows the data that were collected at the three time points: baseline, kindergarten posttest, and grade 1 follow-up.

Table 2.5. Measures and data collection schedule at baseline, kindergarten posttest, and grade 1 follow-up

Data	Use in analysis	Baseline (fall 2008)	Kindergarten posttest (spring 2009)	Grade 1 follow-up (spring 2010)
Student assessments				
Expressive vocabulary	Outcome	✓	✓	✓
Academic knowledge	Outcome	✓	✓	✓
Listening comprehension	Outcome	✓	✓	
Passage comprehension	Outcome	✓		✓
Student characteristics (district administrative records)	Covariates		✓	
School characteristics	Covariates	✓		

Data on student performance for use as covariates in the impact analyses were collected at baseline (fall 2008), before the start of the K-PAVE intervention. Student outcomes were collected again at the end of the kindergarten school year (spring 2009) and one year later (spring 2010), at the end of grade 1. Impacts of K-PAVE at the end of the intervention year were estimated on expressive vocabulary, academic knowledge, and listening comprehension measured at the end of kindergarten. Impacts of K-PAVE one year after the end of the intervention were estimated on expressive vocabulary, academic knowledge, and passage comprehension measured at the end of grade 1. Impacts on instructional practices were assessed only at the end of the intervention year.

The study also addressed exploratory research questions about the impacts of K-PAVE on other kindergarten outcomes not included in the main kindergarten report: components of classroom vocabulary and comprehension support, student lexical diversity, and teacher lexical diversity. These analyses were based on data collected at baseline and at the end of the kindergarten intervention year. Details about these measures, the data collection schedule, the exploratory data analysis, and results are reported in appendix E.

Student assessments

In kindergarten, the primary research question examined the impact on the outcome most directly targeted by K-PAVE—students' expressive vocabulary. 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 was also hypothesized to have a secondary impact on kindergarten students' academic knowledge and listening comprehension. These student outcomes—academic knowledge and listening comprehension—were considered secondary because they extend beyond the primary target of the intervention (table 2.6).[21]

Table 2.6. Student measures, outcome variables, and timing of data collection

Study area	Measure	Outcome variable	Timing	Status in analysis
Vocabulary knowledge	Expressive Vocabulary Test–2 (EVT-2)	Standard score (standardized on norming sample to mean = 100 and standard deviation = 15)	Kindergarten baseline Kindergarten posttest Grade 1 follow-up	Primary confirmatory
Academic knowledge	Woodcock-Johnson III/ Normative Update (WJ-III/NU) academic knowledge subtest (science, humanities, social studies)	*W*-score, an Item Response Theory scale score (all three content areas combined)	Kindergarten baseline Kindergarten posttest Grade 1 follow-up	Secondary confirmatory
Listening comprehension	Kaufman Test of Educational Achievement–II (KTEA-II), listening comprehension subtest	Standard score (standardized on norming sample to mean = 100 and standard deviation = 15)	Kindergarten baseline Kindergarten posttest	Secondary confirmatory
Passage comprehension	Woodcock-Johnson III/ Normative Update (WJ-III/NU) passage comprehension subtest	*W*-score, an Item Response Theory scale score	Kindergarten baseline Grade 1 follow-up	Secondary confirmatory

The follow-up study examined whether the impacts of K-PAVE on kindergarten students were sustained in grade 1, one year after the intervention ended. Three student outcomes were measured in grade 1. Two of the outcomes—expressive vocabulary and academic knowledge—showed impacts at the end of kindergarten. It was hypothesized that greater vocabulary acquisition (the primary outcome) and academic knowledge (a secondary outcome) in kindergarten continued to have positive effects on vocabulary acquisition and academic knowledge in the next school year, even after the end of the intervention. Because there was no

[21] Impacts on classroom instruction in kindergarten (discussed below) were also considered secondary, because they are the intermediate outcomes through which K-PAVE is hypothesized to affect students' expressive vocabulary.

impact on listening comprehension in kindergarten, the study did not investigate sustained impacts in grade 1.

The study also estimated impacts on passage comprehension at the end of grade 1. It was hypothesized that increased vocabulary knowledge at the end of the K-PAVE intervention leads to impacts at the end of grade 1 that go beyond vocabulary to passage comprehension. Impacts on passage comprehension were not examined in kindergarten, because children typically have not yet been exposed to formal reading instruction on decoding text by the end of kindergarten. Because formal reading instruction has begun by grade 1, we conducted a confirmatory analysis of the impact of access to K-PAVE in kindergarten on students' passage comprehension in grade 1.

At each time point, 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 control schools were completed before K-PAVE intervention training (3.9% of students in control schools were assessed one to three weeks later).[22] Posttest assessments for all students were conducted after completion of the 24-week K-PAVE intervention period. Follow-up assessments were conducted in the spring of grade 1, one year after the end of the intervention. Student assessment measures are described below, and student assessment procedures are described in appendix M. The training of data collectors and data quality assurance are described in appendix N.

Expressive vocabulary. Expressive vocabulary is the primary outcome in the impact analysis in both kindergarten and grade 1. The primary measure of vocabulary acquisition was the Expressive Vocabulary Test–2 (EVT-2) (Williams 2007).

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 on which each word passes from receptive into productive use with increasing exposure to the word (see Zareva, Schwanenflugel, and Nikolova 2005). Vermeer (2001) distinguishes between breadth of vocabulary knowledge (the number of words in the lexicon) and depth of such knowledge (how well those words are known). Expressive vocabulary captures both the 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) with only partial word knowledge (Curtis 1987); expressive vocabulary (the ability not only to recognize but also to recall a word) requires a greater depth of knowledge.

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

[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.

statistically significant positive effects on students' expressive vocabulary (Schwanenflugel, et al., 2010). The National Early Literacy Panel (2008) found an average correlation of 0.48 between expressive language comprehension outcomes and later reading performance across 30 studies with random assignment designs. The panel found that measures of receptive vocabulary had lower correlations with both decoding and reading comprehension. Consequently, we concluded that a measure of expressive vocabulary was better aligned with the immediate, primary goals of the K-PAVE program, as well as with the secondary goal of an extended impact on later reading comprehension. We decided not to measure students' receptive vocabulary in addition to their expressive vocabulary because doing so would place additional burden on students, would yield results on a measure both less aligned with intervention goals and highly correlated with expressive vocabulary (the correlation between the PPVT-4 and the EVT-2 is .84; Dunn and Dunn 2007), and would require an adjustment for multiple comparisons in the analysis.

The recently updated version of the Expressive Vocabulary Test (EVT-2) 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. It has been used in criterion studies with other language assessments, such as the Comprehensive Assessment of Spoken Language (CASL) (Carrow-Woolfolk 1999) and the Group Reading Assessment and Diagnostic Evaluation (GRADE) (Williams 2001). It has internal consistency (split half) reliabilities of 0.94–0.95 and test-retest reliability of 0.95 (Williams 2007). Correlations between the EVT-2 and other tests are 0.84 for the Peabody Picture Vocabulary Test–4 (Dunn and Dunn 2007) for children 5–6; 0.68–0.80 for the receptive language, expressive language, and core language scales on the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4) (Semel, Wiig, and Secord 2003) for children 5–8; and 0.59–0.76 for the GRADE total reading score in kindergarten.

EVT-2 raw scores were used to calculate a standard score using student age. The standard scores allowed the performance of students in intervention classrooms to be compared with that of a national sample of children 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. Academic knowledge is a secondary outcome in the impact analyses for both kindergarten and grade 1. The measure of academic knowledge in the study was the Woodcock-Johnson III/Normative Update (WJ-III/NU) Academic Knowledge subtest (Woodcock, McGrew, Schrank, and 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, and Schmitt 2007). The test focuses on background knowledge in science, social studies, and humanities.

Test developers selected a nationally-representative sample of children and adults ages 12 months to 80 years for the WJ-III Normative Update and developed national norms for the WJ-III. Based on student age, nationally-normed standard scores, with a mean of 100 points and a standard deviation of 15 points, were also calculated from raw scores, to describe students' academic knowledge at kindergarten entry. In addition, an Item Response Theory score (*W*-

score) was calculated from the raw score for use in impact analysis models. Test developers used a Rasch single-parameter logistic model to transform raw scores to equal-interval units, which is the *W*-score. Reported technical characteristics of the WJ-III/NU Academic Knowledge test indicate that internal consistency (split half) reliability is 0.92 for five-year-olds and 0.82 for six-year-olds (Woodcock et al. 2007).

Listening comprehension. Listening comprehension is a secondary outcome in the impact analysis in kindergarten. (Because there was no impact of K-PAVE on this outcome in kindergarten, it was not assessed in grade 1). The measure of listening comprehension used in this study is the Kaufman Test of Educational Achievement–II (KTEA-II) Listening Comprehension subtest (Kaufman and Kaufman 2004).

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 best measured by listening to speech. The test of listening comprehension 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 to calculate a single standardized score based on a nationally representative norm group.

Reported technical characteristics of the subtest indicated that the internal consistency (split-half) reliability for this measure in kindergarten is 0.84 (Kaufman and Kaufman 2004). Results from a confirmatory factor analysis with students in grade 1 and higher indicated that the correlation between the oral language factor (listening comprehension and oral language subtests) and the reading factor was 0.91 and that the errors for the listening comprehension subtest and the reading comprehension subtest were correlated. The correlation in grades 1–5 between the KTEA-II Listening Comprehension subtest and the WJ-III Listening Comprehension test was 0.77.

Passage comprehension. A second measure of comprehension in the study was the Passage Comprehension subtest of the WJ-III/NU (Woodcock et al. 2007). This measure is a secondary outcome at the end of grade 1 only. Passage comprehension was not assessed at kindergarten posttest, because kindergarten children typically have not yet had formal reading instruction on decoding text. Nonetheless, the WJ-III/NU Passage Comprehension subtest does include items that are age-appropriate for kindergarten children, and passage comprehension was measured at kindergarten pretest to control for baseline skill and improve the precision of the impact estimate in first grade.

The Passage Comprehension test is a measure of children's comprehension of visually presented material. For kindergarten students, the age-appropriate items begin by assessing symbolic representation using pictorial symbols and move to assessing early print decoding skills. For older children, items include print passages of increasing length to measure early reading comprehension. Specifically, the test begins by asking younger children to match a pictogram to a corresponding picture; it next asks children to match written words to the corresponding picture. It then moves to asking children to verbally complete written sentences with an accompanying. Finally, it asks children to orally complete written sentences without an associated picture. Although the publisher of the instrument indicates that the WJ-III/NU Passage Comprehension assessment is a valid measure for kindergarten students, it does not measure comprehension of extended text until children have mastered the foundational skills of symbolic representation and early decoding.

We measured students' passage comprehension at the end of grade 1 and tested whether there was an impact of K-PAVE in kindergarten on students' passage comprehension in first grade. The kindergarten pretest score on this assessment was included as a covariate in the analysis of K-PAVE impacts on passage comprehension in grade 1. At the beginning of kindergarten, the Passage Comprehension subtest assesses the extent to which students can discriminate visually presented symbols, have knowledge of print conventions and phonological awareness, and have early decoding skills, all of which are important predictors of later text comprehension (National Early Literacy Panel 2008; Whitehurst and Lonigan 2001). Therefore, the kindergarten pretest measure of passage comprehension is an appropriate covariate in the analysis of impacts on passage comprehension at the end of grade 1.

As noted for the Academic Knowledge subscale, national norms for the WJ-III were established by test developers using a nationally-representative sample of children and adults ages 12 months to 80 years. The same scoring methods described for the Academic Knowledge subtest were used to construct nationally-normed standard scores and W-scores for the Passage Comprehension subtest. Reported technical characteristics of the WJ-III/NU Passage Comprehension test indicated that internal consistency (split half) reliability was 0.96 for students 5–7 (Woodcock et al. 2007).

Student characteristics and school characteristics (as covariates)

Data on the characteristics of students and schools were collected and included as covariates in the analysis of impacts on students. Student demographic characteristics (date of birth, gender, race/ethnicity, Individualized Education Plan, and eligibility for free or reduced-price meals) were collected from district administrative records.

Baseline data on school characteristics were gathered from the Mississippi Department of Education, districts, and schools for use as covariates in the analysis. Data were collected on the percentage of students in the school eligible for free or reduced-price meals, the school's Achievement Level Index[23] (a score based on student performance in grades 3 and higher on the

[23] A school's Achievement Level Index is created based on student performance on the Mississippi state accountability test (the Mississippi Curriculum Test [MCT]), administered to all students in grades 3 and higher. The Achievement Level Index corresponds to the School Performance Classification (described above). It is measured on a continuous scale ranging from 100 to 600; the School Performance Classification is a five-level categorical rating (low-performing, underperforming, successful, exemplary, and superior). The Achievement Level Index score is created based on the percentage of students in the school who score basic or higher on the MCT and the percentage of students in the school who score 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."

Mississippi Curriculum Test in 2006/07), the racial/ethnic composition of the school, and the literacy curricula used in kindergarten classrooms.

DESCRIPTION OF ANALYTIC SAMPLE

The analytic sample included 64 schools (30 intervention and 34 control), 128 classrooms (60 intervention and 68 control), and 1,296 students (596 intervention and 700 control). Data collection response rates for all measures and all time points were high (table 2.7).

Table 2.7. Data collection response rates

(percent, except where otherwise indicated)

	Baseline response rate			Posttest response rate			Follow-up response rate		
Level/variable	**Overall**	**Intervention**	**Control**	**Overall**	**Intervention**	**Control**	**Overall**	**Intervention**	**Control**
Schools									
Number	65	31	34	65	31	34	65	31	34
Characteristics	100	100	100	a	a	a	a	a	a
Students									
Number	1,319	619	700	1,319	619	700	1,319	619	700
Characteristics	97.5	95.5	> 99.0	a	a	a	a	a	a
Expressive vocabulary (EVT-2)	96.3	96.0	96.6	93.5	92.2	94.6	87.1	84.3	85.8
Listening comprehension (KTEA-II)	94.4	93.7	95.0	93.4	92.2	94.4	a	a	a
Academic knowledge (WJ-III/NU)	95.1	93.9	96.1	94.3	92.1	94.6	87.1	84.3	85.8
Passage comprehension (WJ-III/NU)	96.3	96.0	96.6	a	a	a	87.0	84.3	85.7

Note: All student demographic data, posttest data, and follow-up data are missing for 1 intervention school, 2 classrooms/teachers, and 23 students because 1 school (2 classrooms) dropped out of the study after baseline data collection.

a. Not applicable because measure was not collected.

Characteristics of teachers and students in intervention and control schools

There were no statistically significant differences in characteristics between teachers (table 2.8) or students (table 2.9) in the intervention and control groups. Nonetheless, all student characteristics were included as covariates in the confirmatory and exploratory analyses of impacts on kindergarten and grade 1 students, because these characteristics are likely related to student performance. Teacher characteristics were not included as covariates in analyses of impacts on students but are presented here to describe the sample. In addition, teacher characteristics were included as covariates in the exploratory analyses of impacts on classroom instruction reported in Appendix E, because teacher characteristics might relate to classroom instructional practices.

Table 2.8. Characteristics of teachers in analytic sample

Characteristic[a]	**Intervention classrooms (*n* = 60)**	**Control classrooms (*n* = 68)**	**Total classrooms (*n* = 128)**	**Test of Difference**[b]
Race/ethnicity, percent				
African American	40.7	48.5	44.9	*t* = –0.81, *p* = .42
White	59.3	51.5	55.1	
Education level, percent				*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, percent				
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
Teaching tenure, years				*t* = 1.05, *p* = .30
Mean	17.0	14.6	15.7	
Standard deviation	12.3	11.3	11.8	
Kindergarten teaching tenure, years				*t* = 0.91, *p* =. 37
Mean	10.9	9.4	10.1	
Standard deviation	9.1	8.7	8.9	

Note: Estimates apply to all intervention and control classrooms, even though data for some classrooms were missing. The distribution of each teacher characteristic for cases with missing data was assumed to be the same as for cases with nonmissing data. Rates of missing data ranged from 0 to 3.3%, depending on the characteristic.

a. Nearly all teachers were female, therefore information on teacher gender is not reported and teacher gender was not included in analyses.

b. A two-level model (teachers within schools) was used to test for baseline differences in teacher characteristics between intervention and control groups. The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates. A multilevel linear model was used to test for differences in dichotomous variables (race/ethnicity [African American/not], early childhood certification [yes/no], and reading certification [yes/no]) and ordinal variables (education level, specified with 3-levels: college=1, some graduate classes=2, and graduate degree=3). Thus, *t*-tests rather than chi-square tests were conducted.

Table 2.9. Characteristics of students in analytic sample

(percent, except where otherwise indicated)

Characteristic	Students in intervention classrooms (*n* = 596)	Students in control classrooms (*n* = 700)	Students in all classrooms (*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/ethnicity				
African American	87.6	82.8	85.0	*t* = 0.81, *p* = .42
Other	12.3	17.2	15.0	
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 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 years, 1.9 months	6 years, 1.8 months	6 years, 1.9 months	*t* = 0.18, *p* = .86
Standard deviation	4.8 months	4.6 months	4.7 months	

Note: Estimates apply to all intervention and control classrooms, even though data for some classrooms were missing. The distribution of each teacher characteristic for cases with missing data was assumed to be the same as for cases with nonmissing data. Rates of missing data ranged from 0.1% to 1.8%, depending on the characteristic.

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

Student outcomes at baseline for intervention and control schools

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), listening comprehension (KTEA-II Listening Comprehension subtest), and passage comprehension (WJ-III/NU Passage Comprehension subtest). For each assessment, test developers established national norms based on nationally-representative samples. Using student age, raw test scores were converted to nationally-normed standard scores with a mean of 100 points and a standard deviation of 15 points. In both intervention and control groups, mean baseline standard scores for students in the sample on expressive vocabulary and academic knowledge were approximately 91 points (see table 2.10). Mean scores in both groups were 9 points below the age norm of 100 points. With a standard deviation of 15 points, a score 9 points below the mean is .60 of a standard deviation below the norm (9/15 = .6). Mean pretest scores in intervention and control groups were slightly lower for listening comprehension, at 88 and 87 points, respectively, or approximately 12 points (or .80 of a standard deviation) below the age norm. On passage comprehension, students in both intervention and controls groups were closer to the age norm, with mean scores of approximately 97 points (i.e., 3 points below the age norm). With a standard deviation of 15 points, scores 3 points below the age norm are approximately 0.2 of a standard deviation below the mean (3/15 = .2).[24]

Table 2.10. Baseline pretest scores on student outcomes

			Test of baseline differences in student outcomes[a]	
Student outcome measure	**Intervention group**	**Control group**	**Estimated difference**	**Test of difference (*p*-value)**
Expressive vocabulary				
Unadjusted pretest mean	91.1	90.7	0.40	.72
Standard deviation	12.3	11.4	1.10	
Number of students	574	676	1,250	
Academic knowledge				
Unadjusted pretest mean	90.5	90.6	–0.05	.96
Standard deviation	11.0	11.3	1.05	
Number of students	574	675	1,249	
Listening comprehension				
Unadjusted pretest mean	88.2	87.0	1.16	.44
Standard deviation	12.8	13.4	1.52	
Number of students	560	665	1,225	
Passage comprehension				
Unadjusted pretest mean	96.9	97.5	–0.59	.47
Standard deviation	10.7	10.5	0.82	
Number of students	574	676	1,250	

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 pretest scores. The model had the same multilevel structure as the model used to test K-PAVE impacts on student outcomes at posttest (see appendix J for impact model specifications). The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates.

[24] Because it is common for kindergarteners to have only minimal if any text decoding skills, it is possible for students to score close to the age-normed mean without decoding any single-word text. Many passage comprehension test items for students entering kindergarten require them to match pictogram representations with illustrated pictures. These initial items test comprehension of symbolic representations but not comprehension of text.

Confirmatory analysis

This section describes the analytic approach used for the confirmatory analysis of follow-up impacts, which examined the impact of the K-PAVE intervention on grade 1 students' expressive vocabulary and grade 1 students' academic knowledge and passage comprehension. (The results of these analyses are reported in chapter 3.) Table 2.11 lists the confirmatory research questions and the level of statistical significance used as the criterion for rejecting the null hypothesis of no intervention impact. As described in the following section, on adjustments for multiple comparisons, we used a more stringent criterion ($p < .025$ instead of $p < .05$) for the secondary confirmatory grade 1 outcomes, in order to reduce the heightened risk of Type I error (false positives) associated with conducting multiple hypothesis tests.

Table 2.11. Criteria for statistical significance of confirmatory research questions

Research question	Criterion for statistical significance
Primary confirmatory research question	
1. Is the impact of K-PAVE on students' expressive vocabulary at the end of kindergarten sustained through the end of grade 1?	$p < .05$
Secondary confirmatory research questions	
2. Is the impact of K-PAVE on students' academic knowledge at the end of kindergarten sustained through the end of grade 1?	$p < .025$
3. Does access to K-PAVE in kindergarten affect students' passage comprehension in grade 1?	$p < .025$

Adjustments for multiple comparisons. In determining whether there is an intervention impact on an outcome, there is some chance that the null hypothesis of no impact is rejected even if there is no true impact (that is, some chance of making a Type I error). To limit the likelihood of false positives to an acceptable level, we set the criterion for rejecting the null hypothesis to $p = .05$ for a single hypothesis test, thereby limiting the probability of a false positive to .05. For the confirmatory analysis of impacts of K-PAVE on students one year after the end of the intervention, there was one primary research question, which was addressed with a single outcome measure—the EVT-2 measured at the end of grade 1. Because there was a single outcome measure, no adjustment for multiple comparisons was required when testing the sustained impact of K-PAVE on expressive vocabulary.

The chance of one or more false positives increases if tests of impact are conducted on more than one outcome. For this reason, it was necessary to make adjustments to reduce the likelihood of false positives when testing impacts on multiple outcomes in a single domain. There were two vocabulary-related outcomes, academic knowledge and passage comprehension, both of which were hypothesized to be affected by K-PAVE instructional practices. One potential approach to reducing the heightened error rate introduced by conducting multiple tests of impacts was to composite related outcomes into a single measure in order to conduct a single test. However, we considered the two secondary outcomes, academic knowledge and passage comprehension, to be distinct constructs and therefore examined the impact of K-PAVE on each outcome separately. Because separate tests for each of the two secondary outcomes were conducted, a Bonferroni correction was applied to protect against the heightened risk of Type I error introduced by conducting these tests. With the Bonferroni correction, a $p < .025$ criterion was used when testing for an impact of K-PAVE on each of the two secondary outcomes at the end of grade 1.

Adjustments were made for multiple comparisons within a single domain but not across substantive domains. We did not combine all outcomes in the study in order to conduct only a single hypothesis test or to apply the Bonferroni correction to all outcomes at once; we did not combine all three student outcomes (expressive vocabulary, academic knowledge, and passage comprehension) in order to conduct a single test or correction. Schochet (2008a) does not recommend adjusting for multiple comparisons across domains, because doing so produces unnecessarily large reductions in statistical power. Consequently, we maintained the substantive distinction between expressive vocabulary as the primary target of the K-PAVE intervention and the other student outcomes as secondary targets and did not adjust for multiple comparisons across the primary and secondary student outcomes.

Statistical power. A statistical power analysis was conducted to determine the sample size target for detecting impacts at the end of first grade, based on assumptions about attrition, the proportion of variance in outcome measures between schools and between classrooms, and the proportion of school-level variation explained by pretest scores and other covariates. (Detailed information on statistical power analyses is presented in appendix F.) The assumptions for the a priori statistical power analysis regarding attrition, intraclass correlation, and school-level R^2 were based on information on these factors from previous evaluation studies with similar populations. The goal in designing the study was to be able to detect impacts on students at the low end of the range of effect sizes found in the previous quasi-experimental study of PAVE, in which effect sizes ranged from 0.20 to 0.43.[25] The lower end of this range was targeted because we expected that longer-term effects at the end of grade 1, following a year of no treatment, were likely to be weaker than effects immediately after the intervention. We attempted to recruit a sufficient number of schools for a minimum detectable effect size of 0.21. However, for the actual sample recruited, the estimated minimum detectable effect size was 0.26. The literature on vocabulary intervention does not include studies of the long-term impacts on vocabulary learning that could serve as a basis for determining the target minimum detectable effect size. Without that evidence, information from the one quasi-experimental evaluation of the intervention was the primary basis for setting the target.

The actual achieved grade 1 minimum detectable effect sizes for expressive vocabulary and academic knowledge were smaller than the estimated minimum detectable effect size of 0.26. As reported in appendix F, actual minimum detectable effect sizes at the end of grade 1 were 0.18 for expressive vocabulary, 0.21 for academic knowledge, and 0.27 for passage

[25] The targeted minimum detectable effect size for classroom instruction outcomes was larger (effect size = 0.51–0.56); larger impacts on classroom instruction were expected, because the intervention is intended to directly impact instructional practices.

comprehension. The actual minimum detectable effect sizes were lower than estimated because of lower sample attrition and higher R^2 values than assumed during the design phase.

Estimating overall impacts of K-PAVE on students. To address each of the confirmatory research questions about the impacts of K-PAVE on all students, we estimated a three-level hierarchical linear model, with school, classroom, and student levels (the model is presented in appendix J). The model provided an estimate of the average impact of the intervention on students across all schools at a given time (for example, at the end of grade 1) and an estimate of the standard error of the impact. The hierarchical linear model was appropriate for this analysis because the study used a multilevel design with students nested within classrooms and classrooms within schools. The multilevel modeling also parsed the variance among students, classrooms, and schools to produce accurate estimated standard errors (Raudenbush and Bryk 2002).

At the student level, the model expresses the student outcome variable (for example, EVT-2 score at grade 1 follow-up) as a function of student baseline test scores (at kindergarten entry) and other student covariates, with residual variation between students within a classroom also represented. (Appendix J expresses this relationship, and those that follow, in mathematical notation.) Residual 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 was in the intervention or the control group. This level also modeled residual variation between schools. The parameter for the intervention variable indicates the impact of the K-PAVE intervention on the specified student outcome. We conducted a *t*-test with a 0.05 level of significance as the criterion for testing the null hypothesis that the intervention effect is zero.[26] A positive and statistically significant parameter estimate means that students in K-PAVE schools scored higher than students in control schools on 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, Deke, Kisker, Puma, and Schochet (2009), because the intervention could affect the standard deviation in the intervention group.[27] Because the range of student test scores within schools was similar to the range of student test scores across all schools, calculating effect sizes using either the overall control group standard deviation (which does not account for the multilevel nature of the data) or within-school standard deviations will yield similar effect sizes. Therefore, we were not concerned that standardized effect sizes did not sufficiently account for the multilevel structure of the data.

[26] For the two secondary confirmatory student outcomes at posttest, academic knowledge and passage comprehension, the Bonferroni correction was applied to reduce the increased Type I error introduced by multiple hypothesis testing. Therefore, a .025 level of significance was the criterion for rejecting the null hypothesis that the intervention impact is zero.

[27] 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.

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

- Score on the student outcome measure at baseline (pretest score).
- Gender.
- Race/ethnicity.
- Eligibility for free or reduced-price meals.
- Special education status (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 the following:

- Reading initiatives (Reading First, a Mississippi state reading initiative, or other).
- Achievement Level Index.
- Percentage of African American students.
- Percentage of students eligible for free or reduced-price meals.
- Locale type (rural, small town, large town/fringe of city).
- Location (within or contiguous with the Mississippi Delta region).

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

Methods for handling missing data. Missing data included student demographic covariates and test scores for students who were not in school for test administration or who had incomplete or incorrectly-administered test batteries. The missing data were imputed separately for intervention and control groups, as described in appendix L.

We imputed missing values using two approaches: (a) single stochastic regression imputation to impute missing scores for outcome variables measured at pretest, posttest, or follow-up and (b) dummy variable adjustment to impute missing student, teacher, and school covariates. For single stochastic regression imputation, 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. For the dummy variable adjustment, all missing cases of a variable were 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 pretest scores.

Table 2.12 summarizes the types of missing data and the approaches for them in the confirmatory analyses of impacts on students in grade 1.

Table 2.12. Missing data in confirmatory analyses

	Percentage missing		
Type of data	**Intervention group**	**Control group**	**Approach for handling in confirmatory analyses**
Student kindergarten pretest score	4.0–6.0	3.0–4.0	Single stochastic regression
Student kindergarten posttest score	4.0	5–6	Single stochastic regression
Student grade 1 follow-up score	12.0	13.0	Single stochastic regression
Student covariates	1.5	1.7	Dummy variable adjustment
School covariates	13.3	8.5	Dummy variable adjustment

Testing assumptions about residuals. The multilevel models used to examine impacts on students assume the normality and homoscedasticity of the residuals. Impact estimates and their standard errors generated from these models are not particularly sensitive to departures from normality. Nonetheless, we examined whether the normality assumption was met by comparing plots of raw residuals for grade 1 outcomes at each level with normal distributions. We found no radical departures from the normality assumption. To evaluate whether the homoscedasticity assumption was met, we examined plots of raw residuals. We found residual variability to be approximately equal at each level of the outcome and treatment indicator.

Sensitivity analysis. The following sensitivity analyses were conducted as part of the confirmatory analysis of impacts of K-PAVE on student outcomes one year after the end of the intervention (see appendix K for details):

- 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 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 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 because of 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, to test the sensitivity of findings to late baseline testing compared with baseline testing conducted earlier in the school year (late baseline testing occurred only in the control group).

- 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 (with an absolute value greater than three).[28]
- Estimating impacts without including students with a raw score of 0 on either a pretest or a follow-up test, to test the sensitivity of the findings to treating a raw score of 0 as a nonresponse rather than 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 refused 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 discusses cases in which the results of the sensitivity analyses deviate from the finding of the main model (the tables in appendix K compare impact estimates from all models). Sources of sensitivity were reported for impacts on individual outcomes, providing transparency about analytic decisions (for example, missing data imputation) or other factors (for example, delayed baseline testing) that may have affected whether an impact was found to be statistically significant.

Exploratory analysis of subgroup differences in impacts of K-PAVE on students

This section describes the analytic approach used for the exploratory analyses of differences in impacts on subgroups of students and schools, for both the kindergarten posttest and the grade 1 follow-up. (The results of the exploratory analyses of subgroup differences in impacts are reported in chapter 4.) Table 2.13 lists the student outcomes that were examined in the subgroup analysis and the timing of measurement.

[28] Studentized residuals are calculated by dividing a raw residual (that is, 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.

Table 2.13. Student outcomes examined in subgroup analysis

Construct	Measure	Variable	Measurement timing
Expressive vocabulary	Expressive Vocabulary Test-2 (EVT-2)	Standard score (standardized on norming sample to mean = 100 and *SD* = 15)	Kindergarten baseline Kindergarten posttest Grade 1 follow-up
Academic knowledge	Woodcock-Johnson III/NU (WJ-III) Academic Knowledge Test	*W*-score, an Item Response Theory–based scale score	Kindergarten baseline Kindergarten posttest Grade 1 follow-up
Listening comprehension	Kaufman Test of Educational Assessment-II (KTEA-II) Listening Comprehension Test	Standard score (standardized on norming sample to mean = 100, standard deviation = 15)	Kindergarten baseline Kindergarten posttest[a]
Passage comprehension	Woodcock-Johnson III/NU (WJ-III) Passage Comprehension Test	*W*-score, an Item Response Theory–based scale score	Kindergarten baseline Grade 1 follow-up[b]

a. Listening comprehension was not measured at the grade 1 follow-up because we did not find a statistically significant impact of K-PAVE on listening comprehension at the end of kindergarten.
b. Impacts on passage comprehension in kindergarten were not examined.

We did not make adjustments for multiple comparisons in the exploratory analyses of subgroup differences in impacts (or in the exploratory analyses of impacts on additional kindergarten outcomes reported in appendix E). Because exploratory analyses are not designed to confirm a priori hypotheses but are intended to investigate K-PAVE impacts further in order to better understand results and generate new hypotheses for future confirmatory analyses, the same stringent criteria for hypothesis testing were not applied. For exploratory analyses, researcher used a .05-level criterion for statistical significance, applied separately to each outcome.

Research has found gender differences in literacy skills at kindergarten entry (see, for example, Ready et al. 2005). In an analysis of data from the 1998/99 Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), which has a nationally representative sample of more than 16,000 kindergarteners, Ready, et al. found that girls enter kindergarten with stronger literacy skills than boys and make greater gains than boys over the course of the school year. The subgroup analysis examined whether the impacts of K-PAVE were more pronounced for boys, helping them catch up with girls; more pronounced for girls, increasing their advantage over boys; or statistically indistinguishable by gender.

As a second exploratory analysis, we asked if the impacts of K-PAVE depended on students' pretest scores. We examined the impacts of K-PAVE for students scoring at least one standard deviation below the age-normed mean on the baseline measure of the outcome and students scoring above that threshold (that is, with scores above one standard deviation below the mean).[29] The subgroup analysis examined whether the impacts of K-PAVE were greater for students entering kindergarten with low pretest scores than for their higher scoring peers, if impacts were greater for higher-scoring students than their lower-scoring peers, or if the impact of K-PAVE did not vary based on pretest score.

Interest in examining the relationship between impacts and the Reading First status of schools stems from the possibility that Reading First schools may already have been using high-quality literacy instruction practices in kindergarten, before K-PAVE was implemented. In Reading First schools, the addition of K-PAVE may have made less of a difference than it did in non-Reading First schools, which could have led to smaller impacts in Reading First schools. Alternatively, it is possible that teachers in Reading First schools have a deeper understanding of the development of children's literacy skills and might therefore have been better able to implement K-PAVE, which would have led to larger K-PAVE impacts in Reading First schools.

Statistical power. Appendix F provides detailed information on the statistical power analysis for the exploratory analysis of subgroup differences in impacts of K-PAVE on student outcomes at the end of kindergarten and grade 1. Information from the confirmatory analysis of impacts of K-PAVE at the end of the kindergarten intervention year was used to estimate minimum detectable differences in impacts for subgroups of students and schools. After completion of the kindergarten study, we had information on the proportion of variance in outcome measures between schools and between classrooms, the amount of school-level variation explained by covariates, and attrition. Relying on the information from the kindergarten study, and assuming 80% power and a *p*-value of .05, we estimated that minimum detectable difference in impacts was 0.24 for subgroups based on gender, 0.40 for subgroups based on pretest score, and 0.34 for subgroups of schools based on Reading First status. The actual achieved minimum detectable differences are reported in appendix F. The ranges were 0.18–0.23 for subgroups of students based on gender, 0.26–0.27 for subgroups of students based on pretest score, and 0.33–0.44 for subgroups of schools based on Reading First status.

Estimating differences in impacts on subgroups of students defined by gender or pretest score. To address the question of differences in impacts for subgroups of students, we added a cross-level interaction to the three-level hierarchical linear model used in the main confirmatory analysis to test for impacts on students overall (the model described in the previous section and specified in appendix J). The modified model included a dummy variable at the student level to indicate membership in a particular subgroup. For example, to examine differences in impacts for subgroups of students defined by gender, we included the same dummy variable for gender (GIRL = 1 for girls and 0 for boys) in the student-level equation (level 1) that was used in the overall impact analysis and included a cross-level interaction of the school-level (level 3) treatment effect with the level 1 student subgroup indicator (GIRL*T, where T = 1 for K-PAVE schools and 0 for control group schools). To examine differences in

[29] We considered using 1.5 standard deviations below the mean (the clinical threshold for identifying children with disabilities) for grouping students. However, examination of the data revealed that too few students scored below this threshold (10% on the Expressive Vocabulary Test-2, 12% on Woodcock-Johnson III/NU academic knowledge test, and 26% on Kaufman Test of Educational Achievement listening comprehension test.). For this reason, one standard deviation below the mean was used as the cutoff for the pretest score groups.

impacts for subgroups of students based on pretest scores, we instead included at level 1 a dummy variable for students with low pretest scores versus other students (LOWENTRY = 1 for students who entered kindergarten with pretest scores one standard deviation or further below the age-normed mean and 0 for students with higher pretest scores), in place of the continuous pretest score variable used in the overall impact analysis, and included a cross-level interaction between the school-level (level 3) treatment effect and the level 1 student subgroup indicator (LOWENTRY*T). A mathematical statement of this model is presented in appendix J.

The parameter estimate for the interaction term in the model indicates the difference in the average impact of K-PAVE for students with low and not low pretest scores as well as for girls and boys. We conducted a *t*-test using a .05-level criterion to reject the null hypothesis that the difference in the average impact for the two subgroups (girls and boys, students with and without low pretest scores) was zero. A statistically significant parameter estimate for the interaction term means that the impact of K-PAVE was estimated to differ based on gender or pretest score. An estimate with a positive value indicates that the impact of K-PAVE was larger for girls than boys (or larger for students with low pretest scores than for students with not low pretest scores); a negative parameter estimate indicates that the impact of K-PAVE was smaller for girls than for boys (or smaller for students with low pretest scores than for students with not low pretest scores). A standardized difference in the impact was calculated for each subgroup comparison by dividing the estimated difference (the parameter estimate for the interaction term in the model) by the standard deviation of the outcome variable in the control group for the full sample. The same student- and school-level covariates that were used in models testing overall impacts of K-PAVE on students were also used in the analysis of differences in impacts for subgroups of students.

Estimating differences in impacts on students in subgroups of schools. To analyze whether the impact of K-PAVE differed in Reading First and non-Reading First schools, we used a similar analytic approach to the one described for subgroups of students. As for student subgroups, we built on the three-level hierarchical linear model for testing impacts on students overall in the confirmatory analysis (described in the section on impacts on students overall and specified in appendix J). In this instance, a dummy variable indicating subgroups of schools (RF = 1 for Reading First schools and 0 for non-Reading First schools) was included in the school-level equation at level 3; this single dummy variable replaced the set of two reading initiative dummy variables that were used in the main impact analysis (i.e., the dummy variable indicating schools with a Mississippi state reading initiative was omitted). The interaction of the school-level subgroup indicator and the school-level treatment indicator was also included in the school-level (level 3) equation (T*RF). A mathematical statement of this model is presented in appendix J.

The estimated parameter for the interaction term indicates the difference in the average impact of K-PAVE for Reading First and non-Reading First schools. We conducted a *t*-test using a .05-level criterion to reject the null hypothesis that the difference in the average impact for Reading First and non-Reading First schools was zero. A statistically significant parameter estimate for the interaction term means that the impact of K-PAVE was estimated to differ based on Reading First status. An estimate with a positive value indicates that the impact of K-PAVE was estimated as larger for Reading First schools than for non-Reading First schools; a negative parameter estimate indicates that the impact of K-PAVE was estimated as smaller for Reading First schools than for non-Reading First schools. A standardized difference in the impact was calculated by dividing the estimated difference (the parameter estimate for the interaction term in the model) by the standard deviation of the outcome variable in the control group for the full sample. The same student- and school-level covariates that were used in models testing overall impacts on students were used in the analysis of differences in impacts on student outcomes for subgroups of schools.

Methods for handling missing data. For the exploratory subgroup analyses, approaches used for handling missing data were the same as those described for confirmatory analyses, with the exception of missing data for student gender.

Student gender. Missing values for student gender were not imputed for the analysis of differences in K-PAVE impacts on boys and girls in kindergarten or grade 1. Although the dummy variable adjustment provided an unbiased impact estimate, the approach can lead to biased estimates of the coefficients for the covariates in a regression model. Because we were interested in coefficients for student gender in the subgroup analysis—as they summarized the differential between girls' and boys' mean outcome scores in the control group and in the treatment group—we did not want the estimates to be biased. For this reason, in the subgroup analysis for boys and girls, models were estimated using listwise deletion for student gender. Student gender was missing for 9 of the 1,296 students (0.7%) in the sample.

Student pretest score. In the confirmatory analyses, missing pretest scores were imputed using single stochastic regression. The imputed values were used in the analysis of subgroups based on pretest score. At baseline, 4%–6% of students in the treatment group and 3%–4% of students in the control group were not tested, depending on the assessment.

School Reading First status. Data on Reading First were available for all schools. Neither imputation of missing data nor listwise deletion of missing cases was required.

Sensitivity analysis. No sensitivity analyses were conducted on subgroup differences in the exploratory impacts of K-PAVE on students at the end of kindergarten and the end of grade 1. The models estimated to examine subgroup differences in impacts differed only trivially from the models for overall impacts, differing only in the addition of a single parameter for estimating the difference in the impact of K-PAVE for the specified subgroups. In all other regards, the models shared the same specification—the same multilevel structure, the same parameters (covariates), and the same assumptions regarding the error terms. Given that the models were nearly identical, we assumed that if the model testing overall impacts was not sensitive to covariate adjustment, missing data imputation, delayed baseline testing, outliers, zero raw scores, or weighting adjustments, the model estimating subgroup differences in impacts would also not be sensitive to those factors. Therefore, we did not consider it necessary to conduct additional sensitivity analyses as part of the subgroup analyses.

3. Confirmatory analysis of impacts on students one year after K-PAVE intervention

The confirmatory analyses in grade 1 examined whether any of the impacts found at the end of kindergarten remained one year later. The analyses investigated impacts in grade 1 on the two outcomes on which there were impacts in kindergarten—expressive vocabulary and academic knowledge. Because there was no impact on students' listening comprehension in kindergarten, the follow-up study did not investigate sustained effects on that outcome in grade 1.

Impacts on passage comprehension were also examined in grade 1. This outcome was not examined in kindergarten, because children typically have not yet been exposed to formal reading instruction on decoding text by the end of kindergarten. Because formal reading instruction has typically begun by grade 1, we conducted a confirmatory analysis of the impact of access to K-PAVE in kindergarten on students' passage comprehension in grade 1.

No evidence was found that the positive and statistically significant impacts of K-PAVE on students' expressive vocabulary and academic knowledge were sustained in grade 1. K-PAVE in kindergarten had no statistically significant impacts on students' expressive vocabulary, academic knowledge, or passage comprehension in grade 1.

This chapter presents the results of analyses that addressed three confirmatory research questions about impacts of the K-PAVE intervention during kindergarten on students' outcomes one year later. The primary confirmatory research question was:

1. Is the impact of K-PAVE on students' expressive vocabulary at the end of kindergarten sustained through the end of grade 1?

The two secondary confirmatory research questions were:

2. Is the impact of K-PAVE on students' academic knowledge at the end of kindergarten sustained through the end of grade 1?

3. Does access to the K-PAVE in kindergarten affect students' passage comprehension in grade 1?

Impacts on students' expressive vocabulary one year after intervention

We began by examining whether the kindergarten impact on the primary outcome — expressive vocabulary (measured by the Expressive Vocabulary Test-2 [EVT-2]) — was sustained through the end of grade 1. The EVT-2 is a one-on-one, normed, 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 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 P.)

We did not find that the impact of K-PAVE on students' expressive vocabulary was sustained in grade 1 (table 3.1).[30] The estimated impact on expressive vocabulary at the end of grade 1 was 0.36 point; the associated effect size of 0.03 was not statistically significant (t = 0.51, p = .61). The 95% confidence interval, which ranged from –1.06 to 1.78 points, included zero. This finding was consistent in all sensitivity analyses (see appendix K). In particular, dropping cases with missing values for the outcome did not yield qualitatively different study findings.

IMPACTS ON STUDENTS' ACADEMIC KNOWLEDGE AND PASSAGE COMPREHENSION ONE YEAR AFTER INTERVENTION

This section presents results of the analysis addressing the two secondary confirmatory research questions. Two subtests of Woodcock-Johnson III/Normative Update (WJ-III/NU) were used to assess students' academic knowledge and passage comprehension. The Academic Knowledge subtest measures students' background knowledge in science, social studies, and humanities. The Passage Comprehension subtest measures students' comprehension of visually presented material. The test begins with pictorial symbols for kindergarten students and with selecting a picture corresponding to a two-word phrase for grade 1 students, and then progresses to comprehension of increasingly longer passages according to the student's ability. For each subtest, an Item Response Theory–scale score (*W*-score) was calculated based on a nationally representative norm group.[31]

[30] For comparison, table 3.1 presents estimated impacts of K-PAVE on student outcomes in kindergarten. If impacts of K-PAVE on the same outcomes measured over consecutive years had been treated as a family of tests requiring an adjustment for multiple comparisons, the threshold for statistical significance would have been lower for each outcome. For expressive vocabulary, on which there was a statistically significant impact in kindergarten (p = .006), the pattern of findings would have been the same, even with the adjustment. For academic knowledge, the impact in kindergarten, which was found to be statistically significant (p-value = .02), would not have met the more stringent threshold for statistical significance had adjustments been made for repeated tests across consecutive years. However, given that the testing of an impact at follow-up was contingent on finding a statistically significant impact in kindergarten, we did not consider it necessary to adjust for multiple tests across consecutive years.

[31] The *W*-scores that were analyzed and reported in table 3.1 are Item Response Theory – scaled and not age normed. Therefore, the nominal units of the scores are not intuitively meaningful. However, the *W*-scores correspond to scores on a standard score scale (with normed mean of 100 and standard deviation of 15). At grade 1 follow-up, standard scale scores on the Academic Knowledge test have a sample mean of 91.7 points (corresponding to a *W*-score of 465.9 points) and a sample standard deviation of 12.4 points (corresponding to 12.3 points on the *W*-score scale). For standard scale Passage Comprehension scores in grade 1, the sample mean was 94.0 points (corresponding to a *W*-score of 452.1 points) and the sample standard deviation 13.9 points (corresponding to 20.5 points on the *W*-score scale).

Table 3.1. Estimated regression-adjusted impact of K-PAVE on student outcomes at end of intervention (kindergarten) and one year later (grade 1)

	Regression-adjusted means					
Focus of research question	**Intervention group**	**Control group**	**Estimated impact (standard error)**	***p*–value**	**95% Confidence interval**	**Effect size**
Expressive vocabulary						
Grade 1	92.0	91.7	0.36 (0.71)	.61	–1.06 to 1.78	0.032[a]
Kindergarten	93.2	91.6	1.60**(0.59)	.006	0.43–2.77	0.141[b]
Academic knowledge						
Grade 1	466.7	465.4	1.23 (0.85)	.14	–0.46 to 2.92	0.098[c]
Kindergarten	456.5	454.5	1.95* (0.85)	.02	0.25–3.65	0.144[d]
Comprehension						
Passage comprehension, Grade 1	451.7	451.2	0.50 (1.69)	.77	–2.89 to 3.88	0.03[e]
Listening comprehension, Kindergarten	90.1	88.7	1.41 (0.88)	.11	–0.35 to 3.17	0.11[f]

* $p < .05$ ** $p < .01$.

Note: The intervention group included 596 students in 60 classrooms in 30 schools; the control group included 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. Kindergarten results are from Goodson et al. 2010.

a. Calculated by dividing estimated impact by standard deviation of control group follow-up EVT-2 score (11.23).

b. Calculated by dividing estimated impact by standard deviation of control group posttest EVT-2 score (11.35).

c. Calculated by dividing estimated impact by standard deviation of control group follow-up WJ-III/NU academic knowledge *W*-score (12.59).

d. Calculated by dividing estimated impact by standard deviation of control group posttest WJ-III/NU academic knowledge *W*-score (13.48).

e. Calculated by dividing estimated impact by standard deviation of control group follow-up WJ-III/NU passage comprehension *W*-score (19.68).

f. Calculated by dividing estimated impact by standard deviation of control group posttest KTEA-II listening comprehension standard score (13.0).

To reduce the heightened chance of Type I error, which is introduced by conducting multiple hypothesis tests, we applied the Bonferroni correction to the individual hypothesis tests for each of the two secondary outcomes. We used a 0.025-level criterion ($p < .025$) when testing for the impact of K-PAVE on the two secondary outcomes. No evidence was found that the impact of K-PAVE on students' academic knowledge was sustained in grade 1. The estimated impact of 1.23 points on academic knowledge (effect size = 0.10 standard deviation) was not statistically significant ($t = 1.46$, $p = .14$). The 95% confidence interval, which ranged from –0.46 to 2.92 points, included zero. This finding was consistent in all sensitivity analyses (see appendix K). In particular, dropping cases with missing values for the outcome did not yield qualitatively different study findings.

No evidence was found of an impact of kindergarten K-PAVE on students' passage comprehension one year after the intervention year. The estimated impact of 0.50 point on passage comprehension (effect size = 0.03) was not statistically significant (*t* = 0.29, *p* = .77). The 95% confidence interval, which ranged from – 2.89 to 3.88 points, included zero. This finding was consistent in all sensitivity analyses (see appendix K).

4. Exploratory analyses of differences in impacts of K-PAVE for subgroups of students and schools

This chapter describes the results of exploratory analyses examining differences in impacts of K-PAVE at the end of kindergarten and at the end of grade 1 for subgroups of students and schools. In contrast to the confirmatory analyses presented in chapter 3, the exploratory research questions were not based on specific a priori hypotheses but investigate the impacts of K-PAVE further in order to determine whether the average impacts differed for girls compared with boys, for students entering kindergarten with low pretest scores compared with students with higher pretest scores, and for students in Reading First schools compared with students in non-Reading First schools.

This chapter examines the following exploratory research questions:

4. Do the impacts of K-PAVE on students' expressive vocabulary, academic knowledge, and listening comprehension at the end of kindergarten and on students' expressive vocabulary, academic knowledge, and passage comprehension at the end of grade 1 vary by student gender and pretest score?
5. Do the impacts of K-PAVE on students' expressive vocabulary, academic knowledge, and listening comprehension at the end of kindergarten and on students' expressive vocabulary, academic knowledge, and passage comprehension at the end of grade 1 differ in Reading First and non-Reading First schools?

We investigated the differences in impact at the end of kindergarten and the end of grade 1 for all primary and secondary student outcomes, including expressive vocabulary, academic knowledge, listening comprehension (kindergarten only), and passage comprehension (grade 1 only). All outcomes were examined regardless of whether or not a statistically significant impact was found for the overall sample, as variation in impact is still possible around a zero mean (or a nonzero mean that is not statistically different from zero).

To summarize findings from these exploratory subgroup analyses, no evidence was found that the impacts of K-PAVE at the end of kindergarten and grade 1 differed, on average, for subgroups of students by gender or pretest score. There was evidence of differential impacts for Reading First and non-Reading First schools on academic knowledge in kindergarten. The overall impact on academic knowledge in kindergarten was statistically significant, with a standardized effect size of 0.14. The estimated impact of K-PAVE on academic knowledge was 0.22 standard deviations in non-Reading First schools; it was not statistically significant in Reading First schools. There was no evidence that the impacts of K-PAVE on other outcomes at the end of kindergarten or for any outcomes at the end of grade 1 differed, on average, in Reading First and non-Reading First schools.

In addition to exploratory subgroup analyses, the results of two other sets of exploratory analyses of kindergarten impacts are reported only in appendix E. First, we examined the impact of K-PAVE on students' and teachers' lexical diversity (an alternative measure of vocabulary production) at the end of the kindergarten intervention year. Although K-PAVE was hypothesized to affect the lexical diversity of students and teachers, the test of this hypothesis was considered exploratory, because the measure itself and the procedures for its measurement are relatively new and there is no strong research on impacts of similar types of interventions on this outcome.

Second, we examined the impact of K-PAVE on each of the four variables that were composited to create the measure of kindergarten classroom vocabulary and comprehension support. In the confirmatory analysis of kindergarten impacts, a positive and statistically significant impact on classroom vocabulary and comprehension support was found at the end of the kindergarten intervention. To better understand what aspects of vocabulary and comprehension support were improved by K-PAVE, we examined the impact of the intervention on each of the four components of the construct: comprehension support provided, higher-order questions asked, and introduction of new vocabulary during book read alouds and introduction of new vocabulary during other instructional time.

ANALYSES OF IMPACT DIFFERENCES FOR SUBGROUPS OF STUDENTS

Student gender

Analysis in this study did not find a statistically significant difference in the estimated impact of K-PAVE by gender on any student outcomes, in kindergarten or in grade 1 (table 4.1). Therefore, the null hypothesis that the impact of K-PAVE is the same for girls and boys could not be rejected for any of the student outcome measures.

Table 4.1. Test of difference in impact of K-PAVE on kindergarten and grade 1 outcomes in girls and boys

Outcome/subgroup	Regression-adjusted means		Estimated impact (standard error)	*p*–value	Effect size
	Intervention group	Control group			
Expressive vocabulary					
Kindergarten					
Overall	93.22	91.62	1.60** (0.59)	.006	0.14
Girls	93.06	91.57	1.48		
Boys	93.45	91.65	1.80		
Difference[a]	–0.40[b]	–0.08	-0.32 (0.73)	.66	–0.03
Grade 1					
Overall	92.02	91.66	0.36 (0.71)	.61	0.032
Girls	91.66	91.16	0.49		
Boys	92.35	92.14	0.21		
Difference	–0.70	–0.98	0.28 (0.79)	.72	0.025

Academic knowledge					
Kindergarten					
Overall	456.48	454.53	1.95* (0.85)	.02	0.14
Girls	456.13	454.70	1.43		
Boys	456.87	454.38	2.49		
Difference	–0.74	0.32	–1.06 (0.85)	.21	–0.08
Grade 1					
Overall	466.65	465.42	1.23 (0.85)	.14	0.10
Girls	466.41	465.19	1.21		
Boys	466.90	465.64	1.26		
Difference	–0.50	–0.45	–0.05 (0.95)	.96	–0.004
Listening comprehension (kindergarten)					
Overall	90.13	88.72	1.41 (0.88)	.11	0.11
Girls	89.66	89.19	0.47		
Boys	90.52	88.16	2.36		
Difference	–0.86	1.03	–1.89 (1.07)	.08	–0.15
Passage comprehension (grade 1)					
Overall	451.72	451.22	0.50 (1.69)	.77	0.03
Girls	454.72	454.59	0.13		
Boys	448.70	447.93	0.77		
Difference	6.02	6.66	–0.64 (2.16)	.77	–0.03

* *p* < .05** *p* < .01.

a. Differences in estimates (i.e., regression-adjusted means or estimated impacts) are calculated by subtracting estimates for boys from estimates for girls.

b. Rounding error is the source of discrepancies between the reported difference (e.g., -0.40) and the difference obtained by subtracting reported estimates for boys from reported estimates for girls (e.g., 93.06 – 93.45 = –0.39).

Expressive vocabulary. In kindergarten the estimated impact on expressive vocabulary for girls was 0.32 point lower than the estimated impact for boys, a difference of 0.03 standard deviation, which was not statistically significant (*t* = –0.44, *p* = .66). The null hypothesis that on average the impact of K-PAVE on expressive vocabulary in kindergarten is the same for boys and girls could therefore not be rejected.

In grade 1 the estimated impact on expressive vocabulary for girls was 0.28 point higher than the average impact for boys, a difference of 0.03 standard deviation, which was not statistically significant (*t* = 0.36, *p* = .72). The null hypothesis that on average the impact of K-PAVE on expressive vocabulary in grade 1 is the same for boys and girls could therefore not be rejected.

Academic knowledge. In kindergarten the estimated impact of K-PAVE on academic knowledge was 1.1 points lower for girls than for boys, a difference of 0.08 standard deviation, which was not statistically significant (*t* = –1.25, *p* = .21). The null hypothesis that on average the impact of K-PAVE on academic knowledge in kindergarten is the same for boys and girls could therefore not be rejected.

In grade 1 estimated impact of K-PAVE on academic knowledge was 0.05 point lower for girls than for boys, a difference of –0.004 standard deviation, which was not statistically significant (*t* = –0.05, *p* = .96). The null hypothesis that on average the impact of K-PAVE on academic knowledge in grade 1 is the same, for boys and girls could therefore not be rejected.

Listening comprehension. In kindergarten the estimated impact of K-PAVE on listening comprehension was 1.9 points lower for girls than for boys, a difference of 0.15 standard deviation, which was not statistically significant (*t* = –1.77, *p* = .08). The null hypothesis that on average the impact of K-PAVE on listening comprehension in kindergarten is the same for boys and girls could therefore not be rejected.

Passage comprehension. In grade 1 the impact of K-PAVE on passage comprehension was 0.64 point lower for girls than for boys, a difference of –0.03 standard deviation, which was not statistically significant (*t* = –0.29, *p* = .77). The null hypothesis that on average the impact of K-PAVE on passage comprehension in grade 1 is the same for boys and girls could therefore not be rejected.

Kindergarten pretest score

We also examined whether the impacts of K-PAVE at the end of kindergarten and the end of grade 1 differed for students who entered kindergarten with low pretest scores on the outcome compared with other students. We tested whether there was a difference in impacts for students scoring at least one standard deviation below the age-normed mean on the baseline measure of the outcome and students scoring above that threshold (that is, scoring higher than one standard deviation below the mean). We did not find a statistically significant difference in the average impact of K-PAVE between the two groups in either kindergarten or grade 1 (table 4.2). The null hypothesis that the impact of K-PAVE on any of the student outcomes is the same for students with low and not low pretest scores could therefore not be rejected.

Table 4.2. Test of difference in impact of K-PAVE on kindergarten and grade 1 outcomes on students with low and not low pretest scores

	Regression-adjusted means				
Outcome/subgroup	**Intervention group**	**Control group**	**Estimated impact (standard error)**	***p*–value**	**Effect size**
Expressive vocabulary					
Kindergarten					
Overall	93.22	91.62	1.60** (0.59)	.006	0.14
Pretest score: LOW[a]	83.12	82.53	0.59		
Pretest score: NOT LOW[b]	97.55	95.61	1.94		
Difference[c]	–14.42[b]	–13.08	–1.35 (1.04)	.20	–0.12
Grade 1					
Overall	92.02	91.66	0.36 (0.71)	.61	0.03
Pretest score: LOW	83.33	83.80	–0.47		
Pretest score: NOT LOW	95.76	95.12	0.64		
Difference	–12.43	–11.33	–1.11 (1.03)	.28	–0.10

Academic knowledge					
Kindergarten					
Overall	456.48	454.53	1.95* (0.85)	.02	0.14
Pretest score: LOW	446.02	443.22	2.79		
Pretest score: NOT LOW	460.58	458.98	1.60		
Difference	–14.56	–15.76	1.20 (1.24)	.33	0.09
Grade 1					
Overall	466.65	465.42	1.23 (0.85)	.14	0.10
Pretest score: LOW	457.59	455.21	2.38		
Pretest score: NOT LOW	470.20	469.52	0.68		
Difference	–12.61	–14.31	1.7 (1.21)	.16	0.13
Listening comprehension (kindergarten)					
Overall	90.13	88.72	1.41 (0.88)	.11	0.11
Pretest score: LOW	83.66	81.59	2.07		
Pretest score: NOT LOW	95.66	94.25	1.41		
Difference	–12.00	–12.66	0.67 (1.26)	.60	0.05
Passage comprehension (grade 1)					
Overall	451.72	451.22	0.50 (1.69)	.77	0.03
Pretest score: LOW	444.46	445.73	–1.26		
Pretest score: NOT LOW	452.59	451.94	0.65		
Difference	–8.13	–6.21	–1.92 (3.24)	.55	–0.10

* $p < .05$ ** $p < .01$

a. LOW is defined as having a pretest score one standard deviation or more below the age-normed mean.

b. NOT LOW is defined as having a pretest score that is not more than one standard deviation below the age-normed mean.

c. Differences in estimates (i.e., regression-adjusted means or estimated impacts) are calculated by subtracting estimates for students with pretest scores that are not low from estimates for students with low pretest scores.

d. Rounding error is the source of discrepancies between the reported difference (e.g., –14.42) and the difference obtained by subtracting reported estimates for boys from reported estimates for girls (e.g., 83.12 – 97.55 = –14.43).

Expressive vocabulary. In kindergarten the estimated impact of K-PAVE on expressive vocabulary for students with low pretest scores was 1.35 points lower than for students with not low pretest scores. The difference (0.12 standard deviation) was not statistically significant ($t = –1.29$, $p = .20$). The null hypothesis that on average the impact of K-PAVE on kindergarten expressive vocabulary is the same for students with low and not low pretest scores could therefore not be rejected.

In grade 1 the estimated impact of K-PAVE on expressive vocabulary for students with low pretest scores was 1.11 points lower than the impact on students with not low pretest scores. The difference of 0.10 standard deviation was not statistically significant ($t = –1.08$, $p = .28$). The null hypothesis that on average the impact of K-PAVE on expressive vocabulary in grade 1 is the same for students with low and not low pretest scores could therefore not be rejected.

Academic knowledge. In kindergarten the estimated impact of K-PAVE on academic knowledge for students with low pretest scores was 1.20 points higher than for students with not low pretest scores. The difference (0.09 standard deviation) was not statistically significant ($t = 0.97$, $p = .33$). The null hypothesis that on average the impact of K-PAVE on kindergarten academic knowledge is the same for students with low and not low pretest scores could therefore not be rejected.

In grade 1 the estimated impact of K-PAVE on academic knowledge was 1.7 points higher for students with low pretest scores than for students with not low pretest scores. The difference (0.13 standard deviation) was not statistically significant (*t* = 1.40, *p* = .16). The null hypothesis that on average the impact of K-PAVE on academic knowledge in grade 1 was the same for students with low and not low pretest scores could therefore not be rejected.

Listening comprehension. In kindergarten the estimated impact of K-PAVE on listening comprehension for students with low pretest scores was 0.67 point higher than the impact for students with not low pretest scores. The difference (0.05 standard deviation) was not statistically significant (*t* = 0.53, *p* = .60). The null hypothesis that on average the impact of K-PAVE on kindergarten listening comprehension is the same for students with low and not low pretest scores could therefore not be rejected.

Passage comprehension. In grade 1 the estimated impact of K-PAVE for students with low pretest scores was 1.92 points lower than the impact for students with not low pretest scores. The difference (0.10 standard deviation) was not statistically significant (*t* = –0.59, *p* = .55). The null hypothesis that on average the impact of K-PAVE on passage comprehension in grade 1 is the same for students with low and not low pretest scores could therefore not be rejected.

ANALYSES OF IMPACT DIFFERENCES FOR SUBGROUPS OF SCHOOLS

No statistically significant difference was found in the average impact of K-PAVE for Reading First and non-Reading First schools for two of the three student outcomes measured in kindergarten (expressive vocabulary and listening comprehension) or for any of the three student outcomes measured in grade 1 (expressive vocabulary, academic knowledge, and passage comprehension) (table 4.3). The null hypothesis that on average the impact of K-PAVE on each of these student outcomes measured at the end of kindergarten and the end of grade 1 is the same in both Reading First schools and non-Reading First schools could therefore not be rejected. For academic knowledge in kindergarten, a statistically significant difference in the impact of K-PAVE on Reading First and non-Reading First schools was found (*t* = –2.02, *p* = .04), suggesting that the impact was larger in non-Reading First schools than in Reading First schools.

Table 4.3. Test of difference between impact of K-PAVE on kindergarten and grade 1 outcomes in Reading First and non-Reading First schools

	Regression-adjusted means				
Outcome/subgroup	**Intervention group**	**Control group**	**Estimated impact (standard error)**	***p*–value**	**Effect size**
Expressive vocabulary					
Kindergarten					
Overall	93.22	91.62	1.60** (0.59)	.006	0.14
Reading First schools	92.66	92.15	0.52		
Non-Reading First schools	93.42	91.43	1.99		
Difference[a]	–0.75[b]	0.72	–1.48 (1.34)	.27	–0.13
Grade 1					
Overall	92.02	91.66	0.36 (0.71)	.61	0.03
Reading First schools	92.51	92.10	0.41		
Non-Reading First schools	91.84	91.50	0.34		
Difference	0.67	0.60	0.07 (1.64)	.97	0.006
Academic knowledge					
Kindergarten					
Overall	456.48	454.53	1.95* (0.85)	.02	0.14
Reading First schools	454.88	455.71	–0.83 (1.61)[c]	.60	–0.06
Non-Reading First schools	457.07	454.07	2.99** (0.97)	.002	0.22
Difference	–2.19	1.64	–3.83*(1.89)	.04	–0.28
Grade 1					
Overall	466.65	465.42	1.23 (0.85)	.14	0.10
Reading First schools	466.37	465.65	0.72		
Non-Reading First schools	466.75	465.33	1.42		
Difference	–0.38	0.32	–0.70 (1.95)	.72	–0.06
Listening comprehension (kindergarten)					
Overall	90.13	88.72	1.41 (0.88)	.11	0.11
Reading First schools	90.68	87.96	2.71		
Non-Reading First schools	89.81	88.96	0.85		
Difference	0.87	–1.00	1.87 (2.03)	.36	0.14
Passage comprehension (grade 1)					
Overall	451.72	451.22	0.50 (1.69)	.77	0.03
Reading First schools	449.63	451.03	–1.40		
Non-Reading First schools	452.48	451.27	1.21		
Difference	–2.84	–0.24	–2.60 (3.89)	.50	–0.13

* $p < .05$ ** $p < .01$.

a. Differences in estimates (i.e., regression-adjusted means or estimated impacts) are calculated by subtracting estimates for boys from estimates for girls.

b. Rounding error is the source of discrepancies between the reported difference (e.g., –0.75) and the difference obtained by subtracting reported estimates for boys from reported estimates for girls (e.g., 92.66 – 93.42 = –0.76).

c. Statistical tests of the impact of K-PAVE on kindergarteners' academic knowledge in Reading First schools and in non-Reading First schools are reported because there was a statistically significant difference in the impact between the two groups of schools.

Expressive vocabulary

In kindergarten the estimated impact of K-PAVE on expressive vocabulary was 1.5 points lower for students in Reading First schools than for students in non-Reading First schools. The difference (0.13 standard deviation) was not statistically significant (*t* = –1.10, *p* = .27). The null hypothesis that on average the impact of K-PAVE on kindergarten expressive vocabulary is the same in both Reading First and non-Reading First schools could therefore not be rejected.

In grade 1 the estimated impact of K-PAVE on expressive vocabulary was 0.07 point higher for students in Reading First schools than for students in non-Reading First schools. The difference (0.006 standard deviation) was not statistically significant (*t* = 0.04, *p* = .97). The null hypothesis that on average the impact of K-PAVE on expressive vocabulary in grade 1 is the same in both Reading First and non-Reading First schools could therefore not be rejected.

Academic knowledge

A statistically significant difference was found in the magnitude of the impact of K-PAVE on kindergarten students' academic knowledge for students in Reading First schools compared with students in non-Reading First schools (*t* = –2.02, *p* = .04). The average estimated impact of K-PAVE for students in Reading First schools was 3.8 points lower than for students in non-Reading First schools, a difference of 0.28 standard deviation. The 95% confidence interval for the estimated difference in the impact was –7.6 to –0.04 points.

The impact of K-PAVE on the two sets of schools was estimated individually. In Reading First schools, the estimated impact on kindergartener's academic knowledge was –0.8 point (effect size = –0.06), which was not statistically significant (*t* = –0.52, *p* = .60). The 95% confidence interval of –4.0 to 2.4 points included zero. In non-Reading First schools, the impact of K-PAVE on kindergarten students' academic knowledge was positive and statistically significant (*t* = 3.08, *p* = .002). In non-Reading First schools in the intervention group, students' academic knowledge scores at the end of kindergarten were an average 2.99 points higher than those of students in non-Reading First schools in the control group. The 95% confidence interval for the estimated impact in non-Reading First schools was 1.0–4.9 points (effect size = 0.22). These findings suggest that there was an impact of K-PAVE on academic knowledge at the end of kindergarten in non-Reading First but not Reading First schools.

In grade 1 the estimated impact of K-PAVE on academic knowledge was 0.70 point lower for students in Reading First schools than for students in non-Reading First schools. The difference (–.06 standard deviation) was not statistically significant (*t* = –0.36, *p* = .72). The null hypothesis that the average impact of K-PAVE on academic knowledge in grade 1 is the same in both Reading First and non-Reading First schools could not therefore be rejected.

Listening comprehension

In kindergarten the estimated impact of K-PAVE on listening comprehension for students in Reading First schools was 1.9 points higher than the impact for students not in Reading First schools. The difference (0.14 standard deviation) was not statistically significant (*t* = 0.92, *p* =

.36). The null hypothesis that the average impact of K-PAVE on listening comprehension in kindergarten is the same in Reading First and non-Reading First schools could therefore not be rejected.

Passage comprehension

In grade 1 the estimated impact of K-PAVE on passage comprehension was 2.6 points lower for students in Reading First schools than for students not in Reading First schools. The difference (–0.13 standard deviation) was not statistically significant (*t* = –0.67, *p* = .50). The null hypothesis that the average impact of K-PAVE on passage comprehension in grade 1 is the same in Reading First and non-Reading First schools could therefore not be rejected.

5. SUMMARY OF FINDINGS AND STUDY LIMITATIONS

This chapter summarizes the findings of the study, describes the main design parameters for the study, and identifies the study's strengths and limitations.

EFFECT OF K-PAVE ON READING-RELATED OUTCOMES IN GRADE 1

The study did not find evidence that the significant impacts of K-PAVE at the end of kindergarten were sustained into grade 1. The difference in measures of expressive vocabulary and academic knowledge of grade 1 students who attended K-PAVE kindergartens the previous year and students in control kindergartens (who did not) was not statistically significant. Moreover, grade 1 students who had attended K-PAVE kindergartens performed no better than students who had not on passage comprehension, a reading-related outcome that was examined at the end of grade 1.

DIFFERENCES IN EFFECTS OF K-PAVE FOR SUBGROUPS OF STUDENTS AND SCHOOLS

The difference in the estimated impact of K-PAVE in kindergarten and grade 1 on girls and boys was not statistically significant for any student outcomes. The null hypothesis that the average impact of K-PAVE is the same for girls and boys could not therefore be rejected for any of the student outcome measures. No statistically significant difference in the average impact of K-PAVE on students with low pretest scores and students with not low pretest scores was found on any student outcome, measured in kindergarten or grade 1. The null hypothesis that the average impact of K-PAVE is the same for students with low and not low pretest scores could not therefore be rejected for any of the student outcomes.

There was a statistically significant difference in the impact of K-PAVE for Reading First and non-Reading First schools on one outcome: academic knowledge measured at the end of kindergarten. The average impact of K-PAVE for students in Reading First schools was 3.8 points lower than for students in non-Reading First schools, a difference of 0.28 standard deviation. In non-Reading First schools, there was a positive and statistically significant impact of K-PAVE on kindergarten students' academic knowledge (effect size = 0.22). In Reading First schools, the impact of K-PAVE on kindergarten academic knowledge was not statistically significant.

There was no statistically significant difference in the impact of K-PAVE for Reading First and non-Reading First schools for other outcomes measured at the end of kindergarten or for any outcomes measured at the end of grade 1. For these outcomes, the null hypothesis that the average impact of K-PAVE is the same for students in Reading First and non-Reading First schools could not be rejected.

STUDY PARAMETERS

The results in this report came from the first randomized controlled trial testing the effectiveness of K-PAVE in enhancing the expressive vocabulary of kindergarten students and sustaining those effects in grade 1. The study employed a cluster randomized design, with approximately 1,300 kindergarten students nested in 64 schools in 35 districts. All 64 schools were in the Mississippi Delta region or surrounding countries, and all volunteered for the study. The study followed students into grade 1 in the same sample of schools. The research design has strong internal validity, based on the randomization procedure used and the sample's low levels of attrition at all levels over the two-year outcome period. The study was well powered to detect impacts on students, and the multilevel modeling used in the analysis appropriately accounted for the nesting of students in classrooms and schools. The study provides statistically unbiased estimates of the impact of K-PAVE in the context it examined—typical implementation conditions in the Mississippi Delta region schools that volunteer for this type of study.

The study tested K-PAVE as implemented under typical rather than optimal conditions. Previous results on fidelity of implementation (see Goodson et al. 2010) indicated 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% implemented 5–7 strategies, and 7% implemented 1–4 strategies. K-PAVE was implemented by regular kindergarten teachers as a supplement to their ongoing literacy instruction, suggesting that schools that want to implement K-PAVE could replicate the study's implementation conditions. As indicated in chapter 1, the developer (Hamilton and Schwanenflugel) has a contract with a publisher to offer K-PAVE and PAVE (which would be called PreK-PAVE). Anyone interested in the program should contact the developer for more information. The impact results reflect the observed effects when the teacher training and support model included 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 its findings, for two main reasons. First, the results apply only to implementation of K-PAVE for one year at the kindergarten level, implemented by teachers who are in their first year of implementation of the program. Second, the findings are not generalizable beyond the Mississippi Delta and surrounding counties. Even within the region, the study was not a random sample of eligible schools but of schools that volunteered to participate (although 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). 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. Compared with schools that declined to participate, volunteer schools tended to have higher percentages of African American students, were more likely to be in small towns, and were less likely to be located in rural areas.

APPENDIX A. K-PAVE MATERIALS PROVIDED TO TEACHERS

The following materials were provided to teachers in intervention schools:

- 48 children's books.
- 1 teacher guide, including
  - 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, including
  - 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 B. SAMPLE WEEKLY UNIT FROM K-PAVE PROGRAM

This appendix provides a sample weekly unit from the K-PAVE program, on transportation.

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	a 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 move with
tire	wheels on a car, truck, or bicycle

Additional words

scooter	something that has two wheels and a tall handle
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 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^3C 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 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 C. List of K-PAVE target words

This appendix lists the 240 K-PAVE target words.

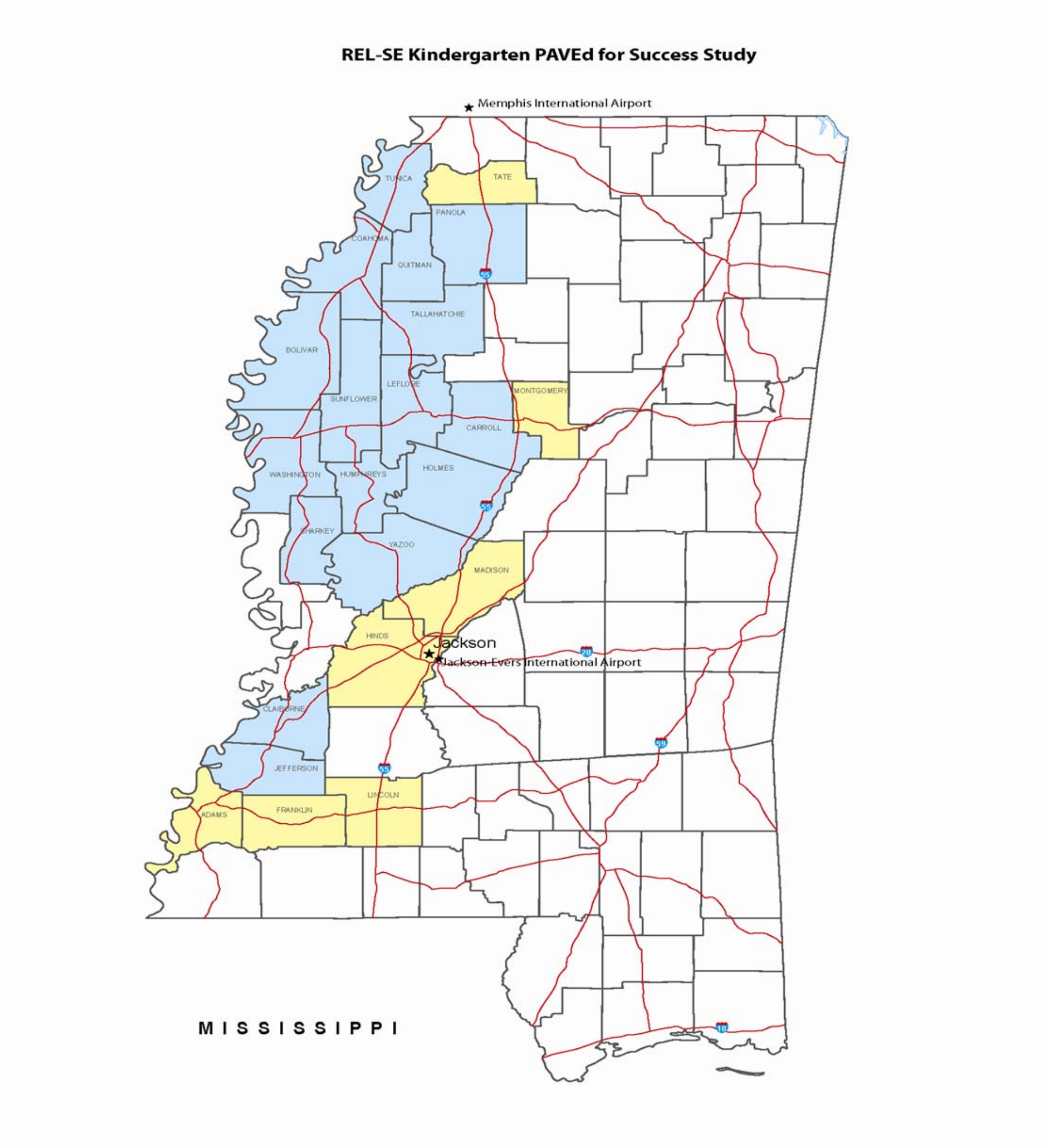
- angry
- antennae
- ants
- aphid
- appear
- arch
- architect
- armor
- artist
- athlete
- atlas
- attract
- backpack
- bacteria
- bait
- baker
- barge
- barn
- battery
- bay
- beetle
- blackberry
- boil
- bored
- boulder
- castle
- cave
- cavity
- cello
- centers
- chick
- chipmunk
- cinnamon
- clarinet
- claws
- cliff
- cloud
- compass
- compost
- conductor
- confused
- container
- cork
- crossing guard
- cub
- cucumber
- delta
- desert
- desk
- disappear
- fireplace
- fishing rod
- flames
- flood
- floss
- flour
- flute
- foal
- forecast
- French horn
- frustrated
- fur
- garden
- globe
- gold
- gravel
- gum
- gymnasium
- hall
- harp
- hatch
- hay
- helicopter
- helmet
- highway
- letter
- library
- lighthouse
- lightning
- liquid
- litter
- locomotive
- lunchbox
- lure
- magnet
- map
- mask
- measuring cup
- melt
- mitten
- mole
- monsoon
- moon
- motor
- mouthwash
- mower
- needle
- nest
- nozzle
- nurse
- plantation
- plastic
- plow
- poison
- pole
- police officer
- praying mantis
- proud
- puddle
- radish
- rake
- recipe
- recycling bin
- repel
- rise
- root
- route
- sailboat
- saliva
- sand dune
- scooter
- seashore
- seeds
- shade
- shed
- star
- steam
- stem
- submarine
- supplies
- sword
- tack
- tambourine
- tarantula
- taxi
- temperature
- thermometer
- thermos
- tide pool
- tire
- track
- trench
- trestle
- trombone
- trumpet
- tunic
- tunnel
- twig
- veterinarian
- volcano

braces
branch
bright
bud
burner
burrow
cable
caboose
cafeteria
calendar
calf
candle
canoe
canyon
cargo
carpenter
disappointed
disgusted
dragon
dragonfly
drill
dusk
earth
earthworm
engineer
equator
eraser
excited
exhaust
exhausted
fawn
feathers
hoe
hook
horseshoe
hose
hydrant
hygienist
icicle
ingredients
iron
judge
junkyard
knight
ladybug
lake
landfill
legend
oar
ocean
orchard
orchestra
owl
paperclip
parents
passageway
pasture
pedal
penny
pickax
picnic
pill bug
planet
plant
shell
shovel
shrimp
shy
sieve
sink
siren
skates
sled
slingshot
sliver
smoke
snowflake
soil
spatula
stall
waist
wasp
waves
web
wheat
whiskers
whistle
wool
wreath
x-ray

APPENDIX D. MISSISSIPPI COUNTIES WITH STUDY SCHOOLS

Map D1. Mississippi counties with study schools, by county

🟦Delta counties
🟨Neighboring counties



APPENDIX E. EXPLORATORY ANALYSES OF KINDERGARTEN IMPACTS ON COMPONENTS OF CLASSROOM VOCABULARY AND COMPREHENSION SUPPORT, STUDENT LEXICAL DIVERSITY, AND TEACHER LEXICAL DIVERSITY

This appendix describes the exploratory analyses of impacts of K-PAVE at the end of the kindergarten intervention year. Exploratory analyses were conducted to examine impacts on components of classroom vocabulary and comprehension support and impacts on alternative measures of student vocabulary development and classroom vocabulary support. Three research questions were addressed. The first deals with components of classroom vocabulary and comprehension support:

- What is the impact of K-PAVE at the end of the intervention on each of the four components of vocabulary and comprehension support in the classroom (introduction of new vocabulary in the classroom during book read alouds, introduction of new vocabulary in the classroom during other instructional time, provision of comprehension support during book read alouds, and use of open-ended questions during book read-alouds)?

The second and third research questions deal with alternative measures of student vocabulary development and classroom vocabulary support:

- What is the impact of K-PAVE on lexical diversity in students' elicited language production, measured at the end of kindergarten?
- What is the impact of K-PAVE on lexical diversity in teachers' naturally occurring language production, measured at the end of the K-PAVE intervention period?

COMPONENTS OF CLASSROOM VOCABULARY AND COMPREHENSION SUPPORT

The study examined the impact of K-PAVE on each of the four individual components of classroom vocabulary and comprehension support:

- Number of words introduced by the teacher or assistant teacher during a book read aloud.
- Average number of words introduced by the teacher or assistant teacher during other instructional time.
- Number of comprehension supports (such as background information, connections to children's experience, clarifications) provided by the teacher or assistant teacher during a book read aloud.
- Number of higher-order questions (asking students to analyze, explain, predict, imagine, make inferences, or generate hypotheses) posed by the teacher or assistant teacher during a book read aloud.

The four variables were combined to create a composite measure of vocabulary and comprehension support, which was examined in the confirmatory analysis of K-PAVE impacts on kindergarten classroom instruction (Goodson et al. 2010). The composite measure was created in order to test the impact of K-PAVE on the overall construct. The K-PAVE intervention is intended to improve teachers' vocabulary and comprehension instructional behaviors, thereby improving students' vocabulary development. Reflecting the intervention goals, we focused on the broad vocabulary and comprehension support construct for the confirmatory analysis. Aggregating the four variables also had the benefit of reducing the number of hypothesis tests conducted in the confirmatory analysis. In addition, the risk of Type I error (that is, false positives) associated with testing the impact of K-PAVE on a single composite measure of vocabulary and comprehension support was lower than the risk associated with testing the impact of K-PAVE on each of the four components of the combined measure.

K-PAVE had a positive and statistically significant impact on the composite measure, with a magnitude of 0.82 standard deviation, equivalent to providing comprehension support 12 more times, asking 3 more higher-order questions, and introducing 3 more vocabulary words during book reading and introducing 3 more words per hour during other instructional times.[32] Analysis of the composite measure cannot indicate the extent to which some or all of the components were responsible for the overall impact on vocabulary and comprehension support. Once a positive and statistically significant impact of K-PAVE was found on the broader composite measure, we undertook an exploratory analysis to examine the impacts of each of the four component variables.

MEASURES AND DATA COLLECTION

Teachers' instructional behavior in both intervention and control classrooms was assessed through direct classroom observation at baseline and at the end of the kindergarten intervention 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.[33] No observations were conducted when study children were in grade 1, because the intervention occurred only at the kindergarten level. Observations were conducted by trained members of the evaluation team who were independent of the K-PAVE intervention and unaware of the school's assignment to the intervention or control group.

Three classroom observation instruments were administered during the intervention year: the Classroom Assessment Scoring System (CLASS K-3) (Pianta, La Paro, and Hamre 2008), the Vocabulary Record, and the Read Aloud Profile-Kindergarten (RAP-K). From these observation instruments, four outcome measures were created for the confirmatory analysis of the impacts of K-PAVE on classroom instruction during the intervention year; results were presented in an earlier report (Goodson et al. 2010). The four classroom instruction outcomes were a composite measure of vocabulary and comprehension support, instructional support,

[32] The estimated impact of 0.82 standard deviation is equivalent to one more word during each 20-minute period observed during instructional time other than the book read aloud. An additional word during every 20 minutes of instructional time suggests a total of three more words per hour during instructional time other than the book read- aloud.

[33] The initial K-PAVE intervention training workshop was staggered over four weeks; thus, 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.

emotional support, and the amount of time spent on literacy instruction in areas other than vocabulary and comprehension. This report includes exploratory analysis of the impacts on the four components of the vocabulary and comprehension support composite (table E1).

Table E1. Classroom measures and vocabulary and comprehension support outcome variables

Study area	Measure	Outcome variable	Status in analysis
Vocabulary support	Vocabulary Record	• Number of words teacher or assistant introduces or asks students to define during book read aloud • Average number of words teacher or assistant introduces or asks students to define during other instructional time	Exploratory
Comprehension support	Read Aloud Profile–Kindergarten	• Number of comprehension supports provided during book read aloud • Number of higher-order questions asked during book read aloud	Exploratory

Two of the four component variables were derived from the Vocabulary Record; the other two were derived from the Read Aloud Profile–Kindergarten. The observation instruments are described below and presented in further detail in appendix Q, which also describes the vocabulary and comprehension support composite.

The Vocabulary Record, which documents the vocabulary support provided by teachers during a book read aloud and during each Classroom Assessment Scoring System (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 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 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 (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 (Smith et al. 2005).

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 that ask students to analyze, explain, predict, imagine, make inferences, or generate hypotheses. Two measures were generated: the number of comprehension supports provided and the number of higher-order questions asked. 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 (that is, occurring zero times) rather than as missing. The RAP-K created for this study is an adaptation of the RAP instrument from the Observation Measures of Language and Literacy Instruction (Goodson, Layzer, Smith, and Rimdzius 2004).

Training of data collectors

Data collectors used a protocol to conduct classroom observations. The SERVE Center at the University of North Carolina at Greensboro recruited classroom observers from local universities and community colleges in Mississippi and through contacts at the Mississippi Department of Education. Data collectors included college students with a background in education, retired teachers, school counselors, and school administrators. Data collectors were independent of the intervention implementation and unaware of the intervention status of a school. As a result, data collection procedures could be maintained as identical in intervention and control schools. (Classroom observation procedures are described in appendix M.)

Classroom observers attended one week of training and had to pass reliability testing before being hired to collect data. Classroom observers were trained to administer the RAP-K and Vocabulary Record instruments by senior Abt Associates staff involved in the development of the instruments or with extensive experience using them. Training for 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 met the required standards. Classroom observers were required to code two video segments of classroom read alouds to demonstrate interrater reliability on the RAP-K and the Vocabulary Record. All videotapes were coded in advance by master coders; observers were required to have at least 80% agreement for all coded video segments. The average rate of agreement with the master codes was 88% on the RAP-K (80%–96%) and 85% on the Vocabulary Record (63%–100%).[34]

Ten of 14 trainees (71%) were certified as reliable for baseline data collection. For posttest data collection, both returning and new data collectors were trained and certified. All four returning observers and five of eight new trainees (63%) were certified as reliable.

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. No measure of interrater reliability was collected during these visits. Within the first week of data collection, trainers held conference calls with data collectors to discuss questions and challenges.

Classroom observation data underwent a thorough quality assurance protocol (described in appendix N).

[34] 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 low incidence of new vocabulary words introduced contributed to a lower than 80% rate of agreement for three classroom observers.

Data collection response rates for classroom vocabulary and comprehension support measures

Table E2 shows response rates for classroom observation measures of vocabulary and comprehension support for both baseline and kindergarten posttest. Response rates for the teacher demographic survey, which was administered at baseline only to collect covariate data on teacher characteristics, are also provided. (The teacher survey is included in appendix R.) The study achieved high response rates at all levels and time points.

Table E2. Data collection response rates for classroom measures of vocabulary and comprehension support and for teacher demographic survey

(percent, except where otherwise indicated)

	Baseline response rate			Kindergarten posttest response rate		
Item	**Overall**	**Intervention group**	**Control group**	**Overall**	**Intervention group**	**Control group**
Number of classrooms	130	62	68	130	62	68
Read Aloud Profile–Kindergarten	100	100	100	97.7	96.8	98.5
Vocabulary Record	100	100	100	98.5	96.8	100
Teacher survey	99.2	98.4	100	a	a	a

Note: One school (two classrooms) dropped out of the study after baseline data collection. As a result, all posttest data are missing for one intervention school with two classrooms/teachers.

a. Not applicable because the measure was not collected.

Vocabulary and comprehension support at baseline for intervention and control schools

Table E3 shows the baseline measures of classroom vocabulary and comprehension support—the composite measure and the four component variables—for intervention and control groups. The two groups did not differ significantly at baseline on these measures: on average teachers provided comprehension support 13–15 times, posed 2 higher-order questions, and introduced 2 vocabulary words during a read aloud and introduced three vocabulary words per hour during other instructional times.[35]

[35] The mean values for the average number of vocabulary words introduced during each 20-minute period observed indicated that teachers introduced an average of one word, which suggests that teachers introduce an average of three words per hour during instructional time other than the book read aloud.

Table E3. Baseline classroom measures

	Raw pretest means		Test of baseline differences in vocabulary and comprehension support outcomes[a]	
Classroom instruction measure	**Intervention group (standard deviation)**	**Control group mean (standard deviation)**	**Estimated difference (standard error)**	**Test of difference (*p*–value)**
Vocabulary and comprehension support composite measure	–0.002 (1.00)	0.002 (1.01)	–0.007 (0.18)	.97
Number of vocabulary words introduced during book read-aloud (Vocabulary Record)	2.08 (2.73)	1.90 (2.12)	0.17 (0.43)	.69
Number of vocabulary words introduced during other times of the school day (Vocabulary Record)	0.88 (1.01)	0.89 (1.04)	–0.008 (0.21)	.97
Number of comprehension supports provided during book read-aloud (Read Aloud Profile–Kindergarten)	13.37 (9.90)	15.15 (14.36)	–1.80 (2.22)	.42
Number of higher-order questions posed during book read-aloud (Read Aloud Profile–Kindergarten)	2.47 (3.20)	2.28 (2.81)	0.18 (0.59)	.76

Note: The sample included all intervention classrooms (60 classes in 30 schools) and all control classrooms (68 classrooms in 34 schools). Vocabulary and comprehension support was measured as a *z*-score with a mean of 0 and a standard deviation of 1. The four component measures of the composite were measured as frequency counts.
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. The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates.

DATA ANALYSIS METHODS

Analytic model

This section describes the analytic approach used for the exploratory analyses of impacts of K-PAVE on the four components of classroom vocabulary and comprehension support. The impacts of the K-PAVE intervention on vocabulary and comprehension support (or any other classroom instruction outcomes) were estimated using a multilevel model to account for the clustering of two classrooms per school. Impacts were estimated controlling for a baseline measure of the vocabulary and comprehension support outcome, as well as teacher and school characteristics. 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 (caused by 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 mathematical terms in appendix J.

If we had sufficient degrees of freedom to include teacher characteristics as classroom-level covariates, we would have done so, as we would expect characteristics of individual teachers to be associated with impacts. Such an association may not show up in the findings when the average of two teachers' characteristics is used instead of individual teachers' characteristics if the association is positive for one teacher and negative for the other. Thus a finding of no association between average teacher characteristics and estimated impacts should not be taken as an indication that no association exists teacher by teacher. On the other hand, a statistically significant association in either direction based on average teacher characteristics implies that for at least one of the two teachers there was an association in that direction.

The same school-level covariates that were included in the models to estimate impacts on students (see appendix O for list of school-level covariates) were included in the models estimating impacts on classroom instruction.

The following teacher characteristics, measured at baseline, were included in the analysis as covariates in the school-level model)[36]

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

Averaging dichotomous teacher characteristics (e.g., African American/White) yields a measure of the percentage of study teachers in each category (0%, 50%, or 100%). For continuous variables (e.g., number of years teaching), the school average indicates the average number of years that study teachers in the school have been teaching. The school averages for characteristics of study teachers are presented in table E4.

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

Table E4. School-level averages for characteristics of study teachers

	Intervention		**Control**	
	(*n*=30 schools)		**(*n*=34 schools)**	
	n	%	*n*	%
Teacher race				
100% white	11	36.7%	12	35.3%
50% African American or missing/50% white	13	43.3%	11	32.4%
100% African American	6	20.0%	11	32.4%
Teacher education				
100% with college degree	4	13.3%	5	14.7%
100% with some graduate courses	0	0.0%	4	11.8%
100% with graduate degree	6	20.0%	5	14.7%
50% college/50% some graduate	9	30.0%	6	17.6%
50% college/50% graduate degree	6	20.0%	9	26.5%
50% some graduate or missing/50% graduate degree or missing	5	16.7%	5	14.7%
Early Childhood Certification				
0% - 50%	13	43.3%	13	38.2%
100%	17	56.7%	21	61.8%
Reading Certification				
0%	23	76.7%	26	76.5%
50% - 100%	7	23.3%	8	23.5%
Teaching tenure, years				
Mean	17.2		14.8	
Standard deviation	9.7		8.7	
Kindergarten teaching tenure, years				
Mean	10.9		9.5	
Standard deviation	7.5		6.1	

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 estimated the impact of K-PAVE on the specified classroom instruction outcome. A *t*-test was conducted with a .05 level of significance as the criterion to test the null hypothesis that the intervention effect is zero. A positive and statistically significant parameter estimate indicated that K-PAVE affects instruction in the desired direction. The magnitude of the parameter indicated the estimated magnitude of the average impact. Effect sizes of classroom impacts were calculated following the approach described in chapter 2 for student impacts (that is, by dividing the estimated impact from the model by the standard deviation of the outcome variable in the control group).

Statistical power

A statistical power analysis was conducted to determine the estimated minimum detectable effect size for impacts on components of vocabulary and comprehension support (see appendix F). Assumptions for the statistical power analysis were based on information from the kindergarten study—specifically, the proportion of variance in the composite variable at each level, the amount of school-level variation explained by pretest score and other covariates, and the actual level of attrition. The estimated minimum detectable effect size was 0.67 standard deviation.

Handling missing data

The approach to handling missing data was the same as that described in chapter 2 for the confirmatory impact analyses. Missing values were imputed using two approaches. Single stochastic regression imputation was used to impute missing values for outcome variables measured at pretest or posttest; dummy variable adjustment was used to impute missing teacher and school covariates (table E5).

Table E5. Missing data in exploratory analysis of components of vocabulary and comprehension support

	Percentage missing		
Type of data	**Intervention**	**Control**	**Approach for handling**
Classroom instruction pretest	0	0	No missing data
Classroom instruction posttest	0	2	Single stochastic regression
Teacher covariates	5.0	1.5	Dummy variable adjustment
School covariates	13.3	8.5	Dummy variable adjustment

SENSITIVITY ANALYSES

Sensitivity analyses were conducted as part of the confirmatory analysis of the composite measure. Missing values for the four component variables (comprehension support during book read-aloud, higher-order questions during read-aloud, vocabulary words during read aloud, and vocabulary words during other times of the school day) were not imputed previously. As part of the exploratory analysis, missing values for the four component variables were imputed using single stochastic regression imputation, as described in the section on methods for handling missing data. A sensitivity model was estimated for each of the four component variables to test the sensitivity of the findings to imputing missing posttest and pretest scores using single stochastic regression imputation compared with casewise deletion (see appendix K).

IMPACTS OF K-PAVE ON COMPONENTS OF VOCABULARY AND COMPREHENSION SUPPORT AT END OF INTERVENTION YEAR (KINDERGARTEN)

The findings suggest that there is a positive and statistically significant impact of K-PAVE on three of the four components of the classroom vocabulary and comprehension support composite measure (table E6). K-PAVE had a positive and statistically significant impact on all three measures of vocabulary and comprehension support during the book read aloud. On average, when reading books aloud to students, teachers in K-PAVE schools provided comprehension support 10.9 more times than teachers in control schools (*t* = 3.27, *p* = .002, effect size = 0.74); asked 2.6 more higher-order questions (*t* = 2.76, *p* = .008, effect size = 0.80); and introduced 2.1 more vocabulary words (*t* = 2.26, *p* = .03, effect size = 0.50). The standardized effect sizes were all smaller than the 0.83 effect size for the vocabulary and comprehension support composite variable. However, the estimated magnitude in nominal units for each of the individual book reading outcomes—11 comprehension supports, 3 higher-order questions, and 2 vocabulary words—were similar to those based on the assumption that the impact on the composite variable is of equal magnitude for each of the individual components of the composite (12 comprehension supports, 3 higher-order questions, and 3 more vocabulary words during the read aloud and 3 more vocabulary words per hour during other instructional times).

Table E6. Estimated regression-adjusted impacts of K-PAVE on components of classroom vocabulary and comprehension support in kindergarten

	Regression-adjusted posttest means					
Focus of research question	**Intervention group**	**Control group**	**Estimated impact[a] (standard error)**	***p*–value**	**95% confidence interval**	**Effect size[b]**
Vocabulary and comprehension support (composite measure)	0.50	–0.23	0.73*** (0.21)	.0009	0.32–1.14	0.83
Vocabulary words during read aloud	5.74	3.64	2.09* (0.92)	.027	0.24–3.94	0.50
Vocabulary words during other times	2.48	1.93	0.54 (0.33)	.109	–0.12–1.21	0.38
Comprehension support	25.71	14.80	10.91** (3.33)	.002	4.25–17.57	0.74
Higher-order questions	5.87	3.26	2.61** (0.95)	.008	0.72–4.51	0.80

*** *p* < .001; ** *p* < .01.

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

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

b. Effect size was 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 the vocabulary and comprehension support composite, 4.12 for vocabulary words during read aloud, 1.42 for vocabulary words during other times, 14.72 for comprehension support, and 3.28 for higher-order questions.

For the one component not measured during book reading (the number of vocabulary words introduced during other instructional times), the estimated impact of 0.5 words per 20 minutes (1.5 more words per hour) had an effect size of 0.38, which was not statistically significant (*t* = 1.63, *p* = .11). This finding suggests that the null hypothesis that there is no difference in the number of vocabulary words introduced by teachers in K-PAVE schools and teachers in control schools during nonreading instructional times could not be rejected.

ALTERNATIVE MEASURES OF STUDENT VOCABULARY DEVELOPMENT AND CLASSROOM VOCABULARY SUPPORT

An additional vocabulary-related outcome, lexical diversity, was measured for both students and teachers. Lexical diversity is a measure of vocabulary use in oral language, derived from the ratio of the number of unique words used in a language sample (often referred to as "types") to the total number of words used (often referred to as "tokens"). Lexical diversity was measured by repeatedly calculating the type-token ratio (TTR) using multiple random samples of the words in the same language transcript (this step corrects for the effects of language sample size). Higher values indicate greater lexical diversity (see Duran, Malvern, Richards, and Chipere 2004; McKee, Malvern, and Richards 2000).

Impacts on lexical diversity were not examined as part of the confirmatory analysis in kindergarten. The measure, for both students and teachers, was considered an exploratory outcome, for the following reasons:

- Both the procedures for measuring lexical diversity and the measure itself are relatively new. Lexical diversity has been reported only in small research studies and primarily in descriptive rather than impact analyses.
- The relationship between lexical diversity and other vocabulary measures is not known.
- Lexical diversity was measured on only a subsample of students: 40% (four students per classroom) were randomly selected for administration of the language production task. The student sample for this analysis was 527 students (245 intervention and 282 control) instead of the full sample of 1,296 students. Based on the power calculations conducted during the design phase, this sample restriction was expected to substantially reduce the statistical power of the test of impacts.

The measure of lexical diversity was included in the study to provide insight into the range of vocabulary students and teachers use in their extended discourse. Measuring the vocabulary used in students' and teachers' language production may enrich the understanding of the confirmatory findings about the impacts of K-PAVE on the expressive vocabulary knowledge demonstrated by students on a standardized assessment and the vocabulary and comprehension support provided in the classroom by teachers.

Measure of student lexical diversity

Student lexical diversity was measured in an Elicited Language Task that captured additional information about students' vocabulary skill not measured by standardized measures of expressive vocabulary, such as the EVT-2. Rather than asking children to name pictured objects and actions, the Elicited Language Task measured the range of words children use in the production of extended discourse. The measure has two parts. First, children were shown photos (for example, a bee on a flower, a child at the doctor, a school bus) and were prompted by the data collector to tell personal narratives (for example, about getting stung by a bee, going to the doctor, breaking a bone, going on a school field trip).[37] Second, children were given a wordless picture book and asked to “read” the story to the data collector, who audio recorded the child’s narrative production.

Based on the systematic transcription and coding of the audio recording of children’s narrative talk during the Elicited Language Task, students’ lexical diversity score was calculated. Student lexical diversity has been found to be moderately correlated with standardized assessments of expressive vocabulary and language development and may identify clinical language problems and the effects of interventions designed to treat them (Goffman and Leonard 2000; Silverman and Ratner 2002).

Measure of teacher lexical diversity

Teachers’ lexical diversity was measured based on a 30-minute sample of an audio recording of teachers’ naturalistic talk during the classroom observation in both intervention and control classrooms. Teachers wore an audio recorder during the entire classroom observation at both baseline and kindergarten posttest. Three 10-minute segments at specified intervals were extracted from the recording and transcribed and coded for analysis. From the transcription, teachers’ lexical diversity score was calculated in the same manner as it was for students.

Teachers’ lexical diversity may be more proximal to the K-PAVE instructional strategies than other classroom outcomes examined in the confirmatory analysis. The specific instructional strategies that teachers are taught to use in K-PAVE—Explicit Vocabulary Instruction, Interactive Book Reading, and extended conversations with children—are all aimed at building students’ vocabulary. It was hypothesized that use of these strategies would result in greater lexical diversity in teachers’ own oral language, which would, in turn, provide a pathway for further enhancing children’s vocabulary knowledge.

Data collection response rates for student and teacher lexical diversity

Response rates for student and teacher lexical diversity were high at both baseline and kindergarten posttest (table E7).

[37] The same data collectors who administered the student assessments described in chapter 2 administered the Elicited Language Task. Data collectors were trained on the procedures for administering the Elicited Language Task as part of the week-long data collection training.

Table E7. Data collection response rates for student and teacher lexical diversity

	Baseline			Kindergarten posttest		
Level/variable	**Overall**	**Intervention group**	**Control group**	**Overall**	**Intervention group**	**Control group**
Students						
Sample size	527	245	282	527	245	282
Percentage response on Elicited Language Task	94.3	94.3	94.3	92.8	93.9	91.2
Teachers						
Sample size	130	62	68	130	62	68
Percentage response on audio recording of teacher talk	93.8	88.7	98.5	93.8	93.4	94.1

Note: One school (two classrooms) dropped out of the study after baseline data collection. As a result, all student demographic data, posttest data, and follow-up data are missing for one intervention school, two classrooms/teachers, and nine students.

Student and teacher lexical diversity at baseline for intervention and control schools

There was no statistically significant difference between intervention and control groups on either student or teacher lexical diversity at baseline (table E8).

Table E8. Baseline student and teacher lexical diversity scores

	Unadjusted pretest means		Test of baseline differences	
Outcome measure	**Intervention group (standard deviation)**	**Control group (standard deviation)**	**Estimated difference (standard error)**	***p*–value**
Student lexical diversity[a]	41.4 (14.0)	43.3 (14.9)	–1.73 (2.20)	.43
Teacher lexical diversity[b]	80.7 (9.01)	81.3 (9.76)	–0.61 (1.94)	.75

Note: The student lexical diversity measure was collected from a random sample of 40% of students in the study (four students per classroom). There were 527 students (245 intervention and 282 control) randomly selected to complete the Elicited Language Task. The teacher lexical diversity measure was collected during the classroom observation, which was conducted in all 60 intervention and 68 control classrooms.

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 lexical diversity (see appendix J for impact model specifications). The test of baseline differences was adjusted for the multilevel structure of the data but was not adjusted for covariates.

b. A two-level model, with classroom and school levels, was used to test for the baseline difference between intervention and control group means in teacher lexical diversity. The test of baseline differences was adjusted for the multilevel structure of the data but not for covariates.

DATA ANALYSIS METHODS

Student lexical diversity

The model used to examine the impact of K-PAVE on students' lexical diversity at the end of kindergarten was the same model used in the confirmatory analyses of impacts on student outcomes (see chapter 2 and appendix J). The three-level hierarchical linear model, with school, classroom, and student levels, provided an estimate of the average impact of the intervention on students' lexical diversity at the end of kindergarten and an estimate of the impact's standard error. At the student level, the model expressed the student lexical diversity as a function of the kindergarten baseline score and student covariates, with residual variation between students within a classroom.[38] 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 to specify whether a school was in the intervention or control group; this level also modeled residual variation between schools. The parameter for the intervention variable indicated the impact of the K-PAVE intervention on the specified student outcome. A *t*-test was conducted with a .05-level of significance as the criterion to test the null hypothesis that the intervention effect was zero. A positive and statistically significant parameter estimate means that students in K-PAVE schools were estimated to have greater lexical diversity than students in control schools 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.

Teacher lexical diversity

The model used to examine the impact of K-PAVE on teachers' lexical diversity at the end of kindergarten was the same model used to examine impacts on components of vocabulary and comprehension support (see appendix J for model specifications).

Statistical power

A statistical power analysis was conducted to determine the estimated minimum detectable effect size for impacts on student lexical diversity and teacher lexical diversity (see appendix F). Assumptions for the statistical power analysis were based on information from the kindergarten study—specifically, the proportion of variance in other student and classroom instruction outcomes at each level, the proportion of school-level variation explained by pretest scores and other covariates, and the actual level of attrition. The estimated minimum detectable effect size for student lexical diversity was 0.29 standard deviation, and the estimated minimum detectable effect size for teacher lexical diversity was 0.48 standard deviation.

Handling missing data

[38] The same student and school covariates used to examine impacts on lexical diversity were used in the confirmatory analysis of impacts on other student outcomes. Covariates are listed in chapter 2 and defined in appendix O.

For measures of lexical diversity, 2–12 percent of data were missing (table E9). The approach to handling missing data was the same as that described in chapter 2 for the confirmatory impact analyses. Missing values were imputed using two approaches: single stochastic regression imputation to impute missing values for outcome variables measured at pretest or posttest and dummy variable adjustment to impute missing teacher and school covariates.

Table E9. Missing data in exploratory analyses of impacts on student lexical diversity and teacher lexical diversity

	Percent missing	
Type of data	**Intervention group**	**Control group**
Student lexical diversity		
Pretest	6	6
Posttest	6	8
Teacher lexical diversity		
Pretest	12	2
Posttest	3	6

Sensitivity analyses

The same sensitivity models conducted in the confirmatory impact analyses were used in the exploratory analysis of lexical diversity (see chapter 2 and appendix K). The models estimated the sensitivity of the findings to covariate adjustment, missing data imputation, delayed baseline testing, outliers, and weighting schools to adjust for the loss of the school that dropped out of the study. No models were required to test the sensitivity of findings to task administration errors or to raw scores of zero, as these situations did not occur for lexical diversity scores.

For exploratory analyses of the impact of K-PAVE on teachers' lexical diversity at the end of the intervention, we conducted the same sensitivity models estimated in the confirmatory analysis of K-PAVE impacts on classroom instruction (see Goodson et al. 2010):

- Estimating a baseline model with two levels (school, classroom/teacher) and no covariates other than the intervention indicator in the level 2 equation and the baseline lexical diversity score in the level 1 equation.
- Estimating impacts without imputing missing values for posttest and pretest lexical diversity 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 lexical diversity score 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 outliers, to test the sensitivity of findings to a few influential cases compared with the exclusion of values with large studentized residuals (with an absolute value greater than three).

- Estimating impacts without weighting schools, to adjust for the loss of one school that dropped out of the study.

Impacts of K-PAVE on students' lexical diversity at the end of the intervention year (kindergarten). The estimated mean difference in students' lexical diversity scores was 0.47 (effect size = 0.04), which was not statistically significant ($t = 0.32$, $p = .75$) (table E10). The 95% confidence interval around the impact estimate, which ranged from –2.46 to 3.39, covered zero. We could not reject the null hypothesis that K-PAVE had no impact on kindergarten students' lexical diversity. These findings remained consistent in the sensitivity analysis (see appendix K).

Table E10. Estimated regression-adjusted impact of K-PAVE on students' lexical diversity at end of intervention year (kindergarten)

	Regression-adjusted means		Estimated impact (standard error)			
Focus of research question	**Intervention group**	**Control group**		***p*–value**	**95% Confidence interval**	**Effect size[a]**
Students' lexical diversity	36.5	36.0	0.47 (1.46)	.75	–2.46 to 3.39	0.04

Note: For the Elicited Language Task, the intervention group included 245 students in 60 classrooms in 30 schools; the control group included 282 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 in kindergarten was calculated by dividing the estimated impact by the standard deviation of the control group posttest lexical diversity score (13.01).

Impacts of K-PAVE on teachers' lexical diversity at the end of the kindergarten intervention year. The estimated mean difference in teachers' lexical diversity scores was 4.0 points, an effect size of 0.34 standard deviation, which was not statistically significant ($t = 1.76$, $p = .08$) (table E11).[39] The 95% confidence interval around the impact estimate, which ranged from –0.56 to 8.66, covered zero. Therefore, we could not conclude that there was an impact of K-PAVE on teachers' lexical diversity.

[39] The magnitude and standard error of the impact estimate remained consistent across all but one of the sensitivity models (see appendix K). In all models but one, the standardized effect size was 0.30–0.34, and the intervention impact was not statistically significant. The findings presented in the text reflect the preponderance of evidence. The exception (a model with just two covariates: intervention status and teachers' baseline lexical diversity score) suggested a larger standardized effect size of 0.48 and a statistically significant impact ($t = 2.74$, $p = .008$). However, not controlling for covariates would mistakenly attribute variation in lexical diversity associated with teacher and school characteristics to the K-PAVE intervention.

Table E11. Estimated regression-adjusted impacts of K-PAVE on teachers' lexical diversity at end of intervention year (kindergarten)

	Regression-adjusted posttest mean		Estimated impact[a] (standard error)			
Focus of research question	**Intervention group**	**Control group**		***p*–value**	**95% confidence interval**	**Effect size[b]**
Teachers' lexical diversity	85.2	80.7	4.05$^{\sim}$ (2.30)	.08	–0.56 to 8.66	0.34

$^{\sim}$*p* < .10.

Note: The intervention group included 60 teachers in 30 schools; the control group included 68 teachers in 34 schools.

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

b. Effect size was 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 standard deviation for the control group was 11.78.

APPENDIX F. STATISTICAL POWER ANALYSIS

This appendix provides details on the statistical power analysis conducted in designing the study. It also reports the study's actual statistical power for detecting impacts at the end of kindergarten, which was determined based on knowledge of the impact results; the estimated minimum detectable effect sizes for follow-up confirmatory and exploratory impacts (also based on knowledge of the kindergarten impact results); and the actual minimum detectable effect sizes for follow-up confirmatory and exploratory impacts. When designing the evaluation of the K-PAVE intervention, we conducted an a priori statistical power analysis to determine the sample size target, which was based on assumptions about attrition, the level of variance in outcome measures between schools and between classrooms, and the proportion of school-level variation explained by pretest scores and other covariates. These assumptions were based on information from previous evaluation studies.

STATISTICAL POWER FOR DETECTING IMPACTS ON STUDENTS

The a priori statistical power analysis suggested that the study would have 80% power to detect impacts of 0.26–0.28 standard deviation or higher for students' expressive vocabulary at the end of grade 1.

A cluster randomized design was used 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):

(F1)

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

where

- $MDES(\overset{\wedge}{\beta}_1\ TREATMENT)$ 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 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 proportion of school-level variance in the outcome explained by covariates.
- R^2_{class} is the proportion of classroom-level variance in the outcome explained by covariates.
- R^2_{child} is the proportion 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.
- 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 of 0.20–0.43, we sought 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 was 0.10–0.15, because the project focused on recruiting schools exclusively within high-poverty communities (see Hedges and Hedberg 2007).[40]
- The proportion of total variation in student outcomes at the classroom level was 0.05. Variation in student outcomes between classrooms was assumed to be relatively low, given that kindergarten classrooms are generally not grouped based on ability.
- The pretest explained 50% of the variation in posttest scores between schools (school-level $R^2 = 0.50$).
- The correlation between any other covariates and the outcome measure, conditional on the value of the pretest measure, was .00 (these covariates did not influence the school-level R^2).[41]
- 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 (that is, the end of grade 1), the student attrition rate was assumed to be approximately 20%.
- A two-tailed test of significance was conducted at the .05 level.
- The desired power to detect effects was 80%.

[40] Based on a compilation of interclass correlations in group-randomized evaluations of student achievement, Hedges and Hedberg (2007) find that the intraclass correlation is approximately 0.19 for low socioeconomic status schools and approximately 0.09 for low-achievement schools.

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

Based on the statistical power analysis, the recruitment target was 60–70 schools, with two classrooms per school and 10 students per classroom.[42] Table F1 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 ranged from 0.25 to 0.29.

Table F1. Minimum detectable effect sizes for student outcomes, by number of schools

School-level intraclass correlation	70 schools	65 schools	60 schools
0.10	0.25	0.26	0.27
0.15	0.27	0.28	0.29

ACTUAL MINIMUM DETECTABLE EFFECT SIZES FOR STUDENT OUTCOMES AT END OF KINDERGARTEN

The actual analytic sample size was 64 schools, 128 classrooms, and 1,296 students (10 students per classroom). Once the kindergarten 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 at the end of the intervention year from the fitted impact models in the kindergarten study. The following equation was used to calculate the actual minimum detectable effect size (MDES) in the sample:

(F2)
$$MDES\,(\hat{\beta_1 treatment}) = Factor\,(\alpha, \beta, df\,) * \frac{S.E.(\hat{\beta_1 treatment}\,)}{SD_{CONTROL}}$$

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*).[43]
- $S.E.(\hat{\beta_1 treatment})$ is the standard error of the estimated treatment impact.

[42] 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 rather 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; we did not want to eliminate those schools from the sample. Based on the assumption of a 20% student attrition rate, eight students per classroom were assumed in the statistical power analysis.

[43] Based on alpha = 0.05, power = 80%, and *df* = 62 (that is, 64 schools – [2 conditions – 1] – 1), the multiplier for the minimum detectable effect size in this study for expressive vocabulary was 2.85 (see Schochet 2008b). For the secondary student outcomes, the alpha level was 0.025, to adjust for the increased risk of Type I error with multiple hypothesis tests; the multiplier for academic knowledge and listening comprehension was 3.13.

- SD_{CONTROL} is the standard deviation of the posttest score in the control group, in the units of the posttest score.

This calculation indicated that, at the end of the intervention year (kindergarten), the minimum detectable effect sizes (in standard deviation units) were 0.144 for the EVT-2, 0.197 for academic knowledge, and 0.212 for listening comprehension (table F2).

Table F2. Estimated and actual minimum detectable effect sizes for student outcomes at end of kindergarten

Outcome	Estimated minimum detectable effect size	Actual minimum detectable effect size
Expressive vocabulary	0.25–0.29	0.14
Academic knowledge	0.25–0.29	0.20
Listening comprehension	0.25–0.29	0.21

Because the minimum detectable effect sizes were lower than anticipated, the assumptions were compared with the observed data (table F3). Appendix J presents a multilevel model estimated to test the impact of K-PAVE on student outcomes, including a list of the student and school-level covariates.

Table F3. Assumed and observed factors related to minimum detectable effect size for student outcomes at the end of kindergarten

		Observed		
Variable	Assumed	EVT-2	Academic knowledge	Listening comprehension
Proportion of variation between schools	0.10–0.15	0.11	0.07	0.10
Proportion of variation between classrooms	0.05	0.01	0.06	0.00
Student attrition rate	0.20	0.07	0.07	0.07
School-level R^2	0.50	0.99	0.76	0.83

The observed between-school variation in posttest scores was within the assumed range; 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 for expressive vocabulary and listening comprehension. Most notable was the difference between the assumed school-level R^2 of 0.50 and the observed school-level R^2 of 0.99 for expressive vocabulary, 0.76 for academic knowledge, and 0.83 for listening comprehension. The school and student covariates in the impact model accounted for two-thirds of the between-school variation in students' posttest academic knowledge scores and nearly all of the between-school variation in students' EVT-2 posttest standard scores.[44]

[44] The student-level covariates in the model were baseline 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); the 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); the percentage of

The higher school-level R^2 and the lower classroom-level variation contributed to the study's power to detect a weaker effect size than had been assumed. Although the study was thought to have 80% power to detect an effect size of 0.26–0.28 or higher, the actual minimum detectable effect size was 0.14 for EVT-2 and 0.20 for academic knowledge. The observed standardized effect sizes of 0.14 for both EVT-2 and academic knowledge were statistically significant (for EVT-2: $t = 2.74$, $p = .006$; for academic knowledge: $t = 2.29$, $p = .022$) (see table F2).

ESTIMATED MINIMUM DETECTABLE EFFECT SIZES FOR STUDENT OUTCOMES AT END OF GRADE 1

To confirm that the study would have sufficient power to detect sustained impacts on students, we conducted a statistical power analysis before undertaking estimating the minimum detectable effect sizes for the follow-up study. The statistical power analysis was not conducted for design purposes, as the study design and the plan to test all students that could be located in study schools at follow-up had already been established. In estimating minimum detectable effect sizes for the confirmatory analysis of impacts at follow-up, we were able to base assumptions on information from the kindergarten data. Specifically, the proportion of variance in outcomes at each level, the proportion of school-level variation explained by pretest score and other covariates, and the actual level of attrition at follow-up informed assumptions for the statistical power analysis. Minimum detectable effect sizes for impacts at follow-up were estimated using equation (F1) and the following assumptions, informed by data from the kindergarten study:

- The proportion of total variation in student outcomes at the school level was 0.11 (observed in the kindergarten study).[45]
- The proportion of total variation in student outcomes at the classroom level was 0.01 (observed in the kindergarten study).[46]
- The pretest and other covariates explained 98% of the variation across schools in posttest scores that is not explained by treatment status (that is, school-level R^2 = 0.98).[47]
- The pretest and other covariates explained 64% of the variation across students in posttest scores that is not explained by treatment status (that is, student-level R^2 = 0.64).
- The pretest and other covariates did not explain any variation across classrooms (that is, classroom-level R^2 = 0).[48]

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

[45] Based on academic knowledge scores in the kindergarten sample, the assumed proportion of total variation at the school level was 0.07 for the secondary confirmatory outcomes.

[46] The assumed proportion of total variation in the secondary confirmatory outcomes at the classroom level was 0.06.

[47] The assumed school-level R^2 was 0.76 for the secondary confirmatory outcomes.

[48] The assumed classroom-level R^2 was 0.03 for the secondary confirmatory outcomes.

- By the end of the study, the student attrition rate was 13% (follow-up data were collected for 87% of the kindergarten analytic sample). The minimum detectable effect size assumes no attrition, as values for missing student outcomes were imputed.
- A two-tailed test of significance conducted at the .05 level for the primary confirmatory outcome.[49]
- There were 62 degrees of freedom.
- The desired power to detect effects was 80%.

Based on these assumptions, the minimum detectable effect size for the impact of K-PAVE on the primary confirmatory outcome, expressive vocabulary, at follow-up was 0.11. The minimum detectable effect size was estimated to be 0.107 with no attrition and 0.112 with 13% attrition.

The estimated minimum detectable effect size for an impact on expressive vocabulary in the follow-up study was lower than the minimum detectable effect size of 0.14 for the kindergarten study. Variation in the number of students per classroom may be one reason why the estimated figure was higher than the actual one in kindergarten. Although there were 10 students in the majority of classrooms, the number of students ranged from 3 to 17. There was an average of 20 students per school in the study; however, the number of students per school ranged from 12 to 29. Because of the variation in the number of students per classroom, we expected an actual minimum detectable effect size for an impact on expressive vocabulary that was slightly larger than the estimated 0.11.

For the secondary confirmatory outcomes, academic knowledge and passage comprehension, the estimated minimum detectable effect size was lower than for expressive vocabulary because of the more stringent criteria set for statistical significance ($p < .025$) to adjust for multiple comparisons (see the section in chapter 2 on adjustments for multiple comparisons in confirmatory impact analyses). Based on the assumptions for the secondary confirmatory student outcomes, we estimated the minimum detectable effect size in the follow-up study to be 0.19, which was comparable to the actual minimum detectable effect size in kindergarten of 0.20.

Actual minimum detectable effect sizes for student outcomes at end of grade 1

Once the follow-up data were collected, we calculated the actual minimum effect sizes using the standard errors of the treatment effects estimated by the confirmatory impact models at follow-up. The actual minimum detectable effect sizes were calculated using equation (F2), assuming power of 80% and an alpha level of 0.05 for expressive vocabulary and 0.025 for the two secondary confirmatory outcomes (academic knowledge and passage comprehension). Table F4 shows the estimated and actual minimum detectable effect sizes for impacts on students at the grade 1 follow-up.

[49] A two-tailed test of significance at the .025 level was used for the two secondary confirmatory outcomes.

Table F4. Estimated and actual minimum detectable effect sizes for impacts on student outcomes at end of grade 1

Outcome	Estimated minimum detectable effect size	Actual minimum detectable effect size
Expressive vocabulary	0.11	0.18
Academic knowledge	0.19	0.21
Passage comprehension	0.19	0.27

Estimated minimum detectable effect sizes for exploratory analyses

We conducted a statistical power analysis before undertaking the exploratory analyses of impacts at the end of kindergarten and the end of grade 1, in order to estimate the minimum detectable effect sizes for impacts on subgroups of students and schools, on kindergarten students' lexical diversity (for a subsample of students), on teachers' lexical diversity, and on components of classroom vocabulary and comprehension support. These statistical power analyses were not conducted for design purposes, as the study design and the plan to test all students that could be located in study schools at follow-up had already been established. In estimating minimum detectable effect sizes for the exploratory analyses, we were able to base assumptions on information from the kindergarten data. Minimum detectable effect sizes for exploratory analysis of impacts were estimated using equation (F1).

Variation in impacts for subgroups of students and schools

To estimate a minimum detectable difference in impacts for subgroups of students or subgroups of schools, we began by estimating the standard error of the impact for each subgroup (based on a set of specified assumptions, noted below). We then used the estimated standard errors for each subgroup to calculate the estimated standard error of the differential in the impact, which could then be used to calculate the estimated minimum detectable difference in impacts. Equation (F1) was used to calculate the minimum detectable effect size for any specified outcome for a subgroup of students or a subgroup of schools, as well as for the full sample. Equation (F3) indicates the part of equation (F1) that estimates the standard error of the impact:

$$(F3) \qquad s.e.(\hat{impact}) = \sqrt{\frac{\sigma^2_{school}(1-R^2_{school})}{sp(1-p)} + \frac{\sigma^2_{class}(1-R^2_{class})}{(sp(1-p))c} + \frac{\sigma^2_{child}(1-R^2_{child})}{(sp(1-p))cn}}\ .$$

Using the estimated standard error of the impact for subgroups derived from equation (F3), we compared the standard error of the difference in the intervention impact for particular subgroups (for example, girls compared with boys), using the following equation:[50]

[50] The equation for the variance of the difference between two random numbers (for example, see Mood, Graybill, and Boes 1950, p. 179) is Variance[X1 – X2] = Variance[X1] + Variance[X2] – 2(Covariance[X1, X2]). We assumed that the variance of the impact estimate for each of the two subgroups (for example, impact on boys and impact on girls) was uncorrelated, making the covariance zero. Hence, the variance of the difference between the two impact estimates was the sum of the variance

(F4) $$s.e.(\hat{\beta}_{A-B}) = \sqrt{[s.e.(\hat{\beta}_A)]^2 + [s.e.(\hat{\beta}_B)]^2}$$

where

- $s.e.(\hat{\beta}_{A-B})$ is the estimated standard error for the difference in the intervention impact for subgroup A compared with the intervention impact for subgroup B (that is, the standard error of the parameter estimate for the interaction between intervention status and the subgroup indicator).
- $s.e.(\hat{\beta}_A)$ is the standard error of the intervention impact estimate for subgroup A (for example, the estimated impact for girls).
- $s.e.(\hat{\beta}_B)$ is the standard error of the intervention impact estimate for subgroup B (for example, the estimated impact for boys).

The estimated standard error for the impact differential between the two subgroups can be substituted into equation (F1) to calculate the minimum detectable difference in the intervention impact for one subgroup compared with another as follows:

(F5) $$MDD(\hat{\beta}_{A-B}) = Factor(\alpha, \beta, df) * \frac{s.e.(\hat{\beta}_{A-B})}{\sigma}.$$

For each exploratory outcome and each subgroup examined in the exploratory analyses, we used sample data to inform assumptions about the proportion of variance at the school, classroom, and student levels and the proportion of variation at each level explained by covariates. We used the same assumptions, drawing on information from the kindergarten study, to estimate minimum detectable differences in impacts for subgroups of students and schools at the end of kindergarten and the end of grade 1. The estimated minimum detectable differences were thus the same for the end of kindergarten and the end of grade 1.

Table F5 lists the assumptions for each parameter in equation (F1) for the exploratory analyses of the intervention impacts for subgroups of students and schools. Analyses of subgroup variation in impacts were conducted for all confirmatory student outcomes in kindergarten and grade 1: (expressive vocabulary, academic knowledge, listening comprehension (kindergarten only), and passage comprehension (grade 1 only); power calculations were based on the primary confirmatory outcome, students' expressive vocabulary in kindergarten.

Results of the power calculations are also presented in table F5. The minimum detectable effect size was larger for subgroups than for the full sample, and the minimum detectable differences, comparing impacts for subgroups, were even larger.

of each impact estimate. Taking the square root of both sides of the equation yielded an estimate of the standard error of the differential between the two impact estimates.

Based on the assumptions in table F1, we estimated that there was 80% power to detect an effect size of 0.17 among boys and 0.17 among girls and a difference between the impacts of 0.24 standard deviation.

The minimum detectable effect size was larger for subgroups of students based on pretest score (students with low baseline scores and students who did not have low baseline scores) than for the full sample or for subgroups based on gender. The larger minimum detectable effect size can be attributed to the lower explanatory power of a model with pretest scores dichotomized into “low” and “not low” rather than specified as a continuous variable. We estimated the minimum detectable effect size to be 0.31 for students with low baseline scores and 0.26 for students with not low baseline scores; the estimated minimum detectable difference between the impacts for these two subgroups was 0.40.

For subgroups of students, the number of schools in each subgroup was the same as the number of schools for the full sample. However, for subgroups of schools—for example, Reading First schools and non-Reading First schools—the subgroups had fewer schools than the full sample. With a smaller number of schools in each subgroup, the minimum detectable effect size was larger than for the full sample. The estimated minimum detectable effect size was 0.33 for Reading First schools and 0.17 for non-Reading First schools. The estimated minimum detectable difference for the impact for Reading First and non-Reading First schools was 0.34 standard deviations.

Table F5. Sample-based assumptions used in power calculations for exploratory subgroup analyses and estimated minimum detectable differences in impacts for subgroups

	Full sample	Gender		Baseline score		Reading First status	
Variable		**Boys**	**Girls**	**Low**	**Not low**	**Reading First**	**No Reading First**
School-level variance (σ^2_{school})	13.80	13.35	15.15	a	10.03	14.85	14.06
Classroom-level variance (σ^2_{class})	1.82	0	1.63	6.43	1.84	5.13	0.75
Student-level variance (σ^2_{child})	114.23	130.1	96.12	75.14	67.59	112.87	112.19
School-level R^2 (R^2_{school})	0.98	0.91	0.94	0.00	0.29	0.88	0.99
Classroom-level R^2 (R^2_{class})	0.00	0.00	0.00	0.00	0.25	0.26	0.00
Student-level R^2 ($R^2_{student}$)	0.64	0.64	0.63	0.07	0.02	0.64	0.65
Number of schools (s)	64	64	64	64	64	17	47
Proportion treatment schools (p)	0.47	0.47	0.47	0.47	0.47	0.47	0.47
Proportion control schools ($1 - p$)	0.53	0.53	0.53	0.53	0.53	0.53	0.53
Number of classrooms per school (c)	2	2	2	2	2	2	2
Number of students per classroom (n)	10	5	5	3	7	10	10
Control group standard deviation (σ)	11.35	12.01	10.42	8.55	9.31	11.06	11.46
Statistical power (β)	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Alpha level (α)	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Degrees of freedom (df)	62	62	62	62	62	15	45
Factor (α, β, df)[b]	2.85	2.85	2.85	2.85	2.85	3.00	2.87
Standard error (impact)	0.04	0.06	0.06	0.11	0.09	0.11	0.06
Minimum detectable effect size	0.11	0.17	0.17	0.31	0.26	0.33	0.17
Standard error (differential)		0.085		0.14		0.12	
Minimum detectable difference		0.24		0.40		0.34	

a. School-level variance could not be estimated for students with baseline scores one or more standard deviations below the age-normed mean.

b. Value is from table 1 of Schochet (2008b, p. 65).

Impacts on students' lexical diversity

Because of the time and cost involved in measuring students' lexical diversity, the measure was collected from only 40% of students randomly selected from the full sample. Using equation (F1), we estimated the minimum detectable effect size for an intervention impact on student lexical diversity at the end of kindergarten, with the sample-based assumptions presented in table F6. We estimated that, using a *p*-value of .05, a power of 80% power would be needed to detect an intervention impact on students' lexical diversity of 0.29 standard deviation.

Table F6. Sample-based assumptions used in power calculations for exploratory analysis of impact of K-PAVE on students' lexical diversity and estimated minimum detectable effect size

Variable	Parameters in equation (F1)	Sample-based assumptions
Proportion of variance at the school level[a]	σ^2_{school}	0.04
Proportion of variance at the classroom level	σ^2_{class}	0.048
Proportion of variance at the student level	σ^2_{child}	0.913
School-level R^2	R^2_{school}	0.7
Classroom-level R^2	R^2_{class}	0
Student-level R^2	$R^2_{student}$	0
Number of schools	*s*	64
Proportion of treatment schools	*p*	0.47
Proportion of control schools	*1–p*	0.53
Number of classrooms/school	*c*	2
Number of students/classroom	*n*	4
Statistical power	*β*	0.80
Alpha level	*α*	0.05
Degrees of freedom	*df*	62
Multiplier	*Factor* (*α, β, df*)	2.85
Standard error (impact)		0.10
Minimum detectable effect size		0.29

a. By substituting the proportion at each level of the total variance in the outcome into equation (F3), we calculated the standard error in standardized units, obviating the need to divide by the control group standard deviation for the outcome (as indicated in equation F1).

Impacts on teachers' lexical diversity

We used a modified version of equation (F1), without the student-level parameters, to calculate the minimum detectable effect size for the impact of K-PAVE on teachers' lexical diversity at the end of the intervention year. Based on the sample-based assumptions in table F7, we estimated that, with a *p* value of .05, 80% power would be needed to detect an intervention impact on teachers' lexical diversity of 0.48 standard deviation.

Table F7. Sample-based assumptions used in power calculations for exploratory analysis of the impact of K-PAVE on teachers' lexical diversity and estimated minimum detectable effect size

Variable	Parameters in equation F1	Sample-based assumptions
Proportion of variance at the school level[a]	σ^2_{school}	0.01
Proportion of variance at the classroom level	σ^2_{class}	0.99
School-level R^2	R^2_{school}	0.7
Classroom-level R^2	R^2_{class}	0
Number of schools	*S*	63
Proportion of treatment schools	*P*	0.47
Proportion of control schools	*1–p*	0.53
Number of classrooms/school	*C*	2
Statistical power	*B*	0.80
Alpha level	*A*	0.05
Degrees of freedom	*df*	62
Multiplier	*Factor (α, β, df)*	2.85
Standard error (impact)		0.17
Minimum detectable effect size		0.48

a. By substituting the proportion at each level of the total variance in the outcome into equation (F3), we calculated the standard error in standardized units, obviating the need to divide by the control group standard deviation for the outcome (as indicated in equation F1).

Impacts on components of vocabulary and comprehension support in the classroom

The last set of exploratory analyses focused on the impact of K-PAVE on each of the four individual variables that were composited to create a measure of vocabulary and comprehension support in the classroom:

- Number of comprehension supports provided during an observed book reading.
- Number of higher-order questions asked during the observed book reading.
- Number of words introduced during the observed book reading.
- Number of words introduced during other instructional time (adjusted for length of observation).

In the confirmatory analysis, K-PAVE had a positive and statistically significant impact of 0.82 standard deviation on the composite measure at the end of the intervention. Based on the standard error of the impact estimated in the confirmatory analysis, we calculated that the study had 80% power to detect an impact of K-PAVE of 0.67 standard deviation for vocabulary and comprehension support (that is, actual minimum detectable effect size = 2.85* standard error [estimated impact]/$\sigma^2_{\text{control group}}$).

Informed by the sample-based assumptions in table F8, we estimated that, with a *p*-value of .05, 80% power was needed to detect intervention impacts of 0.54–0.57 standard deviation for the four components of vocabulary and comprehension support.

Table F8. Sample-based assumptions used in power calculations for exploratory analysis of K-PAVE impact on four components of classroom vocabulary and comprehension support and estimated minimum detectable effect sizes

Variable	Parameter in equation (F1)	Comprehension support (read aloud)	Higher-order questions (read aloud)	Vocabulary (read aloud)	Vocabulary (other times)
School-level variance	$\sigma^2_{\lambda o o \eta \chi \sigma}$	79.9	5.04	6.14	0.77
Classroom variance	$\sigma^2_{\sigma\sigma\alpha\lambda\chi}$	166.75	13.83	11.33	1.56
School-level R^2	R^2_{school}	0.27	0.02	0.09	0.05
Classroom-level R^2	R^2_{class}	0	0	0.08	0.05
Number of schools	s	64	64	64	64
Proportion of treatment	p	0.47	0.47	0.47	0.47
Proportion of control	$1–p$	0.53	0.53	0.53	0.53
Number of classes/schools	c	2	2	2	2
Control standard deviation	σ	14.68	3.3	4.18	1.42
Statistical power	β	0.8	0.8	0.8	0.8
Alpha-level	α	0.05	0.05	0.05	0.05
Degrees of freedom	df	62	62	62	62
Multiplier	$Factor\ (\alpha, \beta, df)$	2.85	2.85	2.85	2.85
Standard error (impact)		0.19	0.2	0.2	0.2
Minimum detectable effect size		0.54	0.57	0.56	0.57

ACTUAL MINIMUM DETECTABLE EFFECT SIZES FOR EXPLORATORY ANALYSES

The actual minimum detectable effect sizes were calculated using equation (F2), assuming power of 80% and an alpha level of 0.05. From each of the exploratory impact models, the estimated standard error for the intervention impact (for main effect models examining impacts on student lexical diversity, teacher lexical diversity, and components of vocabulary and comprehension support) or the estimated standard error for the interaction between intervention status and the subgroup indicator (for models examining differences in impacts for subgroups of students and schools at the end of kindergarten and grade 1) was substituted into equation (F2). Table F9 presents the estimated and actual minimum detectable effect sizes (or minimum detectible differences in effect sizes) for each of the exploratory outcomes.

Table F9. Estimated and actual minimum detectable effect sizes (or minimum detectable differences in impacts) for exploratory analyses

Variable	Estimated minimum detectable effect size	Actual minimum detectable effect size
Differences in impact		
Expressive vocabulary		
Gender–kindergarten	0.24	0.18
Gender–grade 1	0.24	0.20
Pretest–kindergarten	0.39	0.26
Pretest–grade 1	0.39	0.26
Reading First—kindergarten	0.34	0.33
Reading First—grade 1	0.34	0.41
Academic knowledge		
Gender–kindergarten	0.24	0.18
Gender–grade 1	0.24	0.21
Pretest–kindergarten	0.39	0.26
Pretest–grade 1	0.39	0.27
Reading First–kindergarten	0.34	0.39
Reading First–grade 1	0.34	0.43
Listening comprehension		
Gender–kindergarten	0.24	0.23
Pretest–kindergarten	0.39	0.27
Reading First–kindergarten	0.34	0.44
Passage comprehension		
Gender–grade 1	0.24	0.31
Pretest–grade 1	0.39	0.46
Reading First–grade 1	0.34	0.55
Lexical diversity		
Students	0.29	0.32
Teachers	0.48	0.55
Vocabulary and comprehension support components		
Comprehension support in read aloud	0.67	0.63
Higher-order questions in read aloud	0.67	0.82
Vocabulary in read aloud	0.67	0.62
Vocabulary at times other than read aloud	0.67	0.66

APPENDIX G. RANDOM ASSIGNMENT

This appendix describes the random assignment of schools. It details the matching of schools within blocks for random assignment, the process of random assignment, and the concealment of allocation.

MATCHING OF SCHOOLS WITHIN BLOCKS FOR RANDOM ASSIGNMENT

Seventy schools were blocked into three groups based on previous participation in reading initiatives: Reading First (17 schools), the Mississippi state reading initiative (7 schools), or neither Reading First nor a state reading initiative (i.e., local district initiative or no reading initiative) (46 schools). Within these three blocks, schools were matched based on a set of school characteristics (table G1; see table 2.2 in chapter 2 for sample distributions).

Table G1. School characteristics used for matching schools in sample

Measure	Categories
School Performance Classification[a]	Low performing 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 African American students	96–100 81–95 21–80 Less than 20
Locale type	Rural Small town Large town or fringe of city
Location	Within the Delta Contiguous to the Delta

a. Annual classification based on students' performance on the state accountability test (Mississippi Curriculum Test) administered to students in grades 3 and higher.

School characteristics were ordered so that those hypothesized to be more strongly associated with student outcomes were prioritized in the matching process.[51] Within reading initiative blocks, schools were sorted by all five characteristics in the order listed above—from school performance classification to region. Within the three reading initiative blocks, schools were sorted into the three school performance classifications, from lowest performing to highest performing. Within each school performance classification, schools

[51] 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.

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 reflecting 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, within the Delta and contiguous with the Delta.

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; 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 were assigned to the K-PAVE intervention and 35 schools 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. Characteristics that were given lower priority in the matching process (that is, 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. If within the Reading First block, there were six schools classified as low performing or underperforming, they would have been grouped together. These six schools would then have been ordered based on the percentage of students eligible for free or reduced-price meals, four of which had 96% or more students eligible. In two of those schools, 96% or more of all students were African American. By the time the last two characteristics—locale type and location—could be factored into the matching process in this example, there would be only two Reading First schools that could be 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 type and location, these two schools could not be sorted any further; they would already be matched with each other. One school would be randomly assigned to the intervention condition, regardless of the values for locale type and location.

PROCESS OF RANDOM ASSIGNMENT: SEQUENCE GENERATION

Random assignment was conducted by the evaluation team. This team was independent of the intervention team, which was responsible for intervention training and support, and 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 or control condition was conducted using SAS computer software (version 9.2 for Windows). 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 and Moore 1982). Specifically, the following SAS code was used:

```
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. Specifying zero as the seed, as done here, means that the computer clock initializes the stream. Including a multiplier (2, in this case) changes the length of the interval; 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, yielding an integer (1 or 2) with a 50% 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.

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. The assignment of schools on the ordered list thus alternated between each condition until all schools within a block were assigned to either the intervention or control condition.

Table G2 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 G2. Random assignment for a hypothetical list of Reading First schools, ordered based on school characteristics

School ID	School Performance Classification	Percent eligible for free or reduced-price meals	Percent African American	Locale type	Location	Treatment
1	Low performing/ underperforming	96–100	96–100	Rural	Within the Delta	1
2	Low performing/ underperforming	96–100	96–100	Small town	Contiguous to the Delta	0
3	Low performing/ underperforming	90–95	81–95	Rural	Within the Delta	1
4	Low performing/ underperforming	70–89	81–95	Small town	Within the Delta	0
5	Low performing/ underperforming	70–89	21–80	Large town	Contiguous to the Delta	1
6	Successful	96–100	81–95	Rural	Within the Delta	0
7	Successful	90–95	96–100	Rural	Contiguous to the Delta	1
8	Successful	90–95	81–95	Small town	Within the Delta	0
9	Successful	< 70	96–100	Large town	Within the Delta	1
10	Exemplary/superior	90–95	81–95	Rural	Within the Delta	0
11	Exemplary/superior	90–95	21–80	Large town	Within the Delta	1
12	Exemplary/superior	< 70	21–80	Small town	Within the Delta	0
13	—	96–100	96–100	Rural	Contiguous to the Delta	1

— is not applicable

CONCEALMENT OF ALLOCATION

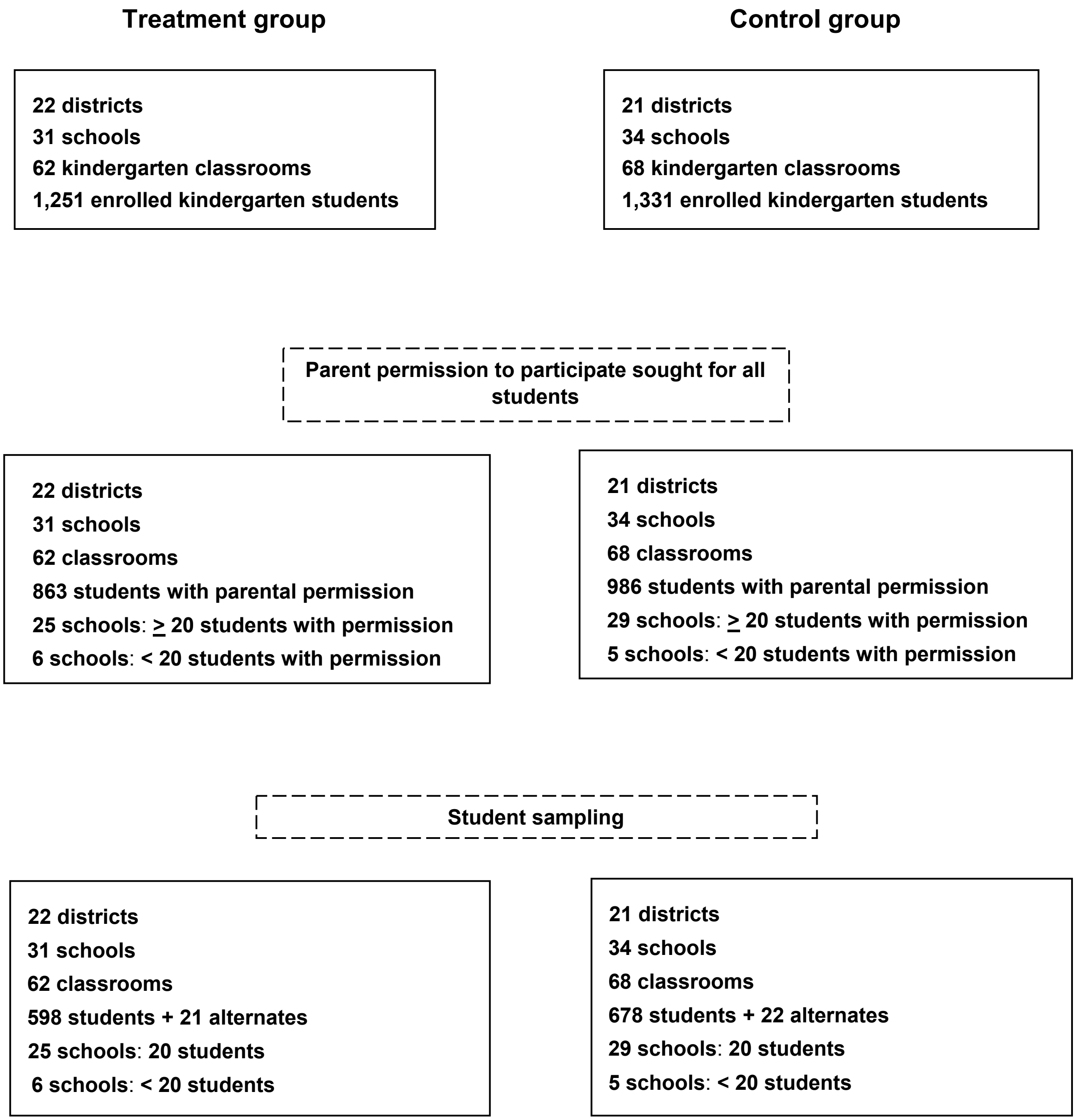
For random assignment to be successful, the process by which study units are randomly allocated to the intervention or control condition must be concealed from both the study participants and the evaluators as the allocation takes place (Forder, Gebski, and Keech 2005). If either study participants or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

In this study, allocation was concealed from both evaluators and study participants during the randomization process. The process for blocking schools based on reading initiative and ordering schools within blocks based on school characteristics was defined before initiating the random assignment process and was not influenced by 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% chance of assignment to either the intervention or control condition. The evaluator had no control over the allocation of schools apart from initiation randomization by the computer.

For study participants, the decision to enroll in the study was not influenced by random assignment. Districts, schools, and teachers were not notified of random assignment status until all required paperwork for study participation was submitted. Concealing the random assignment status from schools until enrollment was completed guarded against the introduction of selection bias as part of the study participation process.

APPENDIX H. RECRUITMENT AND RANDOM SELECTION OF THE STUDENT SAMPLE

This appendix graphically shows the recruitment and random selection of the student sample for this study.



Treatment group

Control group

22 districts
31 schools
62 kindergarten classrooms
1,251 enrolled kindergarten students

21 districts
34 schools
68 kindergarten classrooms
1,331 enrolled kindergarten students

Parent permission to participate sought for all students

22 districts
31 schools
62 classrooms
863 students with parental permission
25 schools: ≥ 20 students with permission
6 schools: < 20 students with permission

21 districts
34 schools
68 classrooms
986 students with parental permission
29 schools: ≥ 20 students with permission
5 schools: < 20 students with permission

Student sampling

22 districts
31 schools
62 classrooms
598 students + 21 alternates
25 schools: 20 students
6 schools: < 20 students

21 districts
34 schools
68 classrooms
678 students + 22 alternates
29 schools: 20 students
5 schools: < 20 students

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

Students missing the baseline assessment did not differ from students who were not missing the baseline assessment, in terms of gender, eligibility for free or reduced-price meals, having an Individualized Education Plan, or age (table I1). A greater percentage of students who completed baseline assessments than those who were missing baseline assessments were African American ($t = 2.49$, $p = .007$).

Table I1. Characteristics of students missing and not missing baseline assessment

(percent, except where otherwise indicated)

Characteristic	**Missing baseline assessment (n = 46)**	**Not missing baseline assessment (n = 1,250)**	**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	92.1	92.9	$t = 0.37$, $p = .71$
Not eligible	7.9	7.1	
Has Individualized Education Plan			
Yes	17.5	7.9	$t = 1.90$, $p = .06$
No	82.5	92.1	
Age at posttest			
Mean	6 years, 2.8 months	6 years, 1.9 months	$t = 1.08$, $p = .28$
Standard Deviation	4.8 months	4.7 months	

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 the 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 [having an Individualized Education Plan]), 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). A chi-square test, which is usually used to test group differences in categorical variables, would not take into account the multilevel structure of the data.

Appendix J. Model specifications

This appendix describes four models used to estimate various impacts.

Three-level model used to estimate overall impacts on students

To model the overall impact of the K-PAVE intervention on students, we estimated a hierarchical linear model. This model provided an estimate of the average impact of the intervention on students across all schools at a given time (for example, at the end of kindergarten or the end of grade 1) as well as an estimate of the standard error of this impact. In the evaluation data, students were nested within classrooms, and classrooms were nested within schools. Therefore, a three-level hierarchical linear model was specified, with students nested within classrooms within schools. The multilevel modeling also parsed 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 same model was used for both confirmatory and exploratory questions about the impact of K-PAVE on students overall rather than for subgroups. This model was used to address the three confirmatory research questions (questions 1–3) about the impact of K-PAVE on student outcomes one year after the end of the intervention and one exploratory research question about the impact of K-PAVE on student lexical diversity in kindergarten (question 7).

The model used to test impacts on students overall 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 resulted in estimates that represented 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, schools that either 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, equation is

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(pre_{ijk} - \overline{pre}) + \beta_{2jk}(female_{ijk} - \overline{female}) \\ &+ \beta_{3jk}(StudentIEP_{ijk} - \overline{StudentIEP}) \\ &+ \beta_{4jk}(FreeLunch_{ijk} - \overline{FreeLunch}) \\ &+ \beta_{5jk}(AfricanAmerican_{ijk} - \overline{AfricanAmerican}) + \varepsilon_{ijk} \end{aligned} \tag{J1}$$

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 receives special education services (has an Individualized Education Plan) and 0 otherwise (centered at the grand mean, $\overline{StudentIEP}$).
- $FreeLunch_{ijk}$ 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 $\overline{FreeLunch}$).
- $AfricanAmerican_{ijk}$ 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.
- $\beta_{1jk} - \beta_{5jk}$ are regression coefficients indicating the effect of each student-level covariate on the outcome measure Y_{ijk}
- 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, equations are

(J2) $\beta_{0jk} = \pi_{00k} + r_{jk}$

(J3) $\beta_{1jk} = \pi_{10k}$

(J4) $\beta_{2jk} = \pi_{20k}$

(J5) $\beta_{3jk} = \pi_{30k}$

(J6) $\beta_{4jk} = \pi_{40k}$

(J7) $\beta_{5jk} = \pi_{50k}$

where

- π_{00k} is the covariate-adjusted mean value of the outcome measure for school *k*.

- 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).
- $\pi_{10k} - \pi_{50k}$ are regression coefficients indicating the average effect in school *k* of each student-level covariate on the outcome measure, Y_{ijk}.

The level 3, or school-level, equations are

(J8)

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

(J9) $\pi_{10k} = \gamma_{100}$

(J10) $\pi_{20k} = \gamma_{200}$

(J11) $\pi_{30k} = \gamma_{300}$

(J12) $\pi_{40k} = \gamma_{400}$

(J13) $\pi_{50k} = \gamma_{500}$

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 the treatment status dummy variable that takes 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 that takes 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 that takes 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 for the k^{th} school (centered at the grand mean, $\overline{AchLvlIndex}$).[52]
- $PctAfrAm_k$ 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 that takes the value 1 if the *k*th school is in a small town and 0 otherwise (centered at the grand mean, $\overline{SmallTown}$).
- $LrgeTown_k$ is a dummy variable that takes 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, $\overline{LrgeTown}$).[53]
- $Delta_k$ is a dummy variable that takes the value 1 if the *k*th school is in the Delta and 0 otherwise (centered at the grand mean, $\overline{Delta}$).
- γ_{002} - γ_{009} are regression coefficients indicating the effect of each school-level covariate on the covariate-adjusted mean value of the outcome measure.
- 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).
- $\gamma_{100} - \gamma_{500}$ are regression coefficients indicating the average effect in school *k* of each student-level covariate on the outcome measure, Y_{ijk}.

The parameter γ_{001} 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 average treatment effect is 0, using a .05-level criterion.[54] A positive and statistically significant estimate of γ_{001} indicates that there is compelling scientific evidence that the K-PAVE intervention improves student vocabulary, academic knowledge, passage comprehension, or lexical diversity. The magnitude of γ_{001} estimates the magnitude of the impact: school participation

[52] 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 (the Mississippi Curriculum Test), which is administered to all students in grades 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 scoring 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."

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

[54] For the two secondary confirmatory outcomes at the end of grade 1, academic knowledge and passage comprehension, a .025-level criterion was used to reject the null hypothesis to reduce the increased likelihood of Type I error that occurs with multiple hypothesis testing.

in K-PAVE is estimated to have, on average, a γ_{001} point effect on student scores 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 et al. (2009), because the intervention might affect the standard deviation in the treatment group. The effect size is $\frac{\hat{\gamma}_{001}}{S_c}$, where S_c is the standard deviation of the outcome measure in the control group.

THREE-LEVEL MODEL USED TO ESTIMATE DIFFERENCES IN IMPACTS FOR SUBGROUPS OF STUDENTS

To address the exploratory question of variation in impacts for subgroups of students at the end of kindergarten and the end of grade 1 (research question 4), we added a cross-level interaction to the three-level hierarchical model specified above. The modified model included a dummy variable to indicate subgroups of students (for example, GIRL = 1 for girls and 0 for boys) at level 1 and a cross-level interaction of the level 3 interaction indicator with the level 1 student subgroup indicator (for example, GIRL*T, where T = 1 for K-PAVE schools and 0 for control group schools).

To estimate the differential in impacts on subgroups of students, we modified the model specified above for testing impacts on students overall as follows:

(J14)

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(girl_{ijk}) + \beta_{2jk}(pre_{ijk} - \overline{pre}) + \beta_{3jk}(StudentIEP_{ijk} - \overline{StudentIEP}) \\ &\quad + \beta_{4jk}(FreeLunch_{ijk} - \overline{FreeLunch}) \\ &\quad + \beta_{5jk}(AfricanAmerican_{ijk} - \overline{AfricanAmerican}) + \varepsilon_{ijk} \end{aligned}$$

where all of the terms in equation (J14) are the same as in equation (J1), except that the subgroup indicator—$girl_{ijk}$, in this case—is not grand mean centered. The variable $girl_{ijk}$ has a value of 1 if the *ith* student in the *jth* classroom in the *kth* school is a girl and 0 if the *ith* student in the *jth* classroom in the *kth* school is a boy. Without grand mean centering, the interpretation of two parameters changes as follows:

- β_{0jk} *is* the covariate-adjusted mean value of the outcome measure for boys in classroom *j* in the *kth* school.

- β_{1jk} is the covariate-adjusted differential between girls and boys in the mean value of the outcome measure for classroom *j* in the *kth* school.

The classroom-level equations were not changed. However, the interpretation of two of the terms in classroom-level equations (J2) and (J3) changed as follows:

- π_{00k} is the covariate-adjusted mean value of the outcome measure for boys in school *k*.

- π_{10k} is the covariate-adjusted differential between girls and boys in the mean value of the outcome measure in school *k*.

One of the school-level equations (J9) above changed as follows:

(J15) $\pi_{10k} = \gamma_{100} + \gamma_{101}(T_k)$.

The interpretation of the terms in (J15) differed from those in equation (J9); the interpretation of two of the terms in equation (J8) changed as follows:

- γ_{000} is the covariate-adjusted mean value of the outcome measure for boys across control schools.
- γ_{100} is the covariate-adjusted differential between girls and boys in the 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) for boys.
- γ_{101} is the mean difference between girls and boys in the covariate-adjusted main effect of treatment.

Substituting the level 2 equations (J2) – (J7) and the level 3 equations (J8), (J10) – (J13), and (J15) into the level 1 equation (J14) yields the following combined model:[55]

(J16) $Y_{ijk} = \gamma_{000} + \gamma_{001}(T_k) + \gamma_{100}(girl_{ijk}) + \gamma_{101}(girl_{ijk})(T_k) ... + \varepsilon_{ijk} + r_{0jk} + v_{00k}$.

The parameter γ_{001} indicates the average impact of the intervention for boys on the specified student outcome; γ_{101} indicates the differential in the average impact of K-PAVE for girls compared with boys. Therefore, the estimated average impact of the intervention for girls is the sum of γ_{001} and γ_{101}.

We conducted a *t*-test to test the null hypothesis that the differential in the average impact of K-PAVE for girls and boys is zero (that is, γ_{101} = 0). A statistically significant estimate of γ_{101} indicates that there is compelling evidence, at the 5% level, that the K-PAVE intervention has a differential impact on girls and boys. An estimate with a positive value indicates that the impact of K-PAVE is larger for girls than for boys; a negative value indicates that the impact of K-PAVE is smaller for girls than for boys. The magnitude of γ_{101} estimates the magnitude of the differential in the impact for girls compared with boys.

[55] Covariates other than intervention status and gender are not shown in equation (J16) in order to simplify the illustration to parameters of interest.

We calculated an effect size for the estimated differential in the impact (γ_{101}) by dividing the parameter estimate from the model by the standard deviation of the outcome variable in the control group for the full sample. Control group standard deviations were used, as recommended in Burghardt et al. (2009), because the intervention could affect the standard deviation in the intervention group.

The same analytic approach (that is, model specification, model estimation, and effect size calculation) outlined for gender was used to examine whether the impacts of K-PAVE varied based on students' performance on the outcome measure at kindergarten entry (that is, the pretest score). The same model specified above (to examine if impacts vary for boys and girls) was used to examine if impacts varied for students based on low baseline scores. The student-level equation (equation J1) is specified as follows:

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(LOWENTRY_{ijk}) + \beta_{2jk}(girl_{ijk} - \overline{girl}) \\ &\quad + \beta_{3jk}(StudentIEP_{ijk} - \overline{StudentIEP}) \\ &\quad + \beta_{4jk}(FreeLunch_{ijk} - \overline{FreeLunch}) \\ &\quad + \beta_{5jk}(AfricanAmerican_{ijk} - \overline{AfricanAmerican}) + \varepsilon_{ijk}\,. \end{aligned} \tag{J17}$$

Instead of including a continuous variable for student pretest score, the model included a dummy variable at level 1 (LOWENTRY) to indicate whether students entered kindergarten with a low score on the outcome measure (LOWENTRY = 1 if students' pretest score was one standard deviation or more below the age-normed mean and 0 if higher). As specified above for boys and girls, the model included a cross-level interaction of the level 3 intervention status variable with the level 1 indicator of low pretest score (that is, LOWENTRY*T, where T = 1 for K-PAVE schools and 0 for control group schools). Substituting the level 3 and level 2 equations into the level 1 equation shown in equation (J17) yielded the following combined model:[56]

$$Y_{ijk} = \gamma_{000} + \gamma_{001}(T_k) + \gamma_{100}(LOWENTRY_{ijk}) + \gamma_{101}(LOWENTRY_{ijk})(T_k) ... + \varepsilon_{ijk} + r_{0jk} + \upsilon_{00k}\,. \tag{J18}$$

We conducted a *t*-test to test the null hypothesis that the differential in the average impact of K-PAVE for students entering kindergarten below grade level and those who did not was zero (that is, $\gamma_{101} = 0$) and used a .05-level criterion to reject the null hypothesis.

[56] School and student covariates other than LOWENTRY and treatment status are not shown in this model, in order to facilitate the illustration of the interaction between pretest score and the treatment condition.

THREE-LEVEL MODEL USED TO ESTIMATE DIFFERENCES IN IMPACTS IN STUDENT OUTCOMES FOR SUBGROUPS OF SCHOOLS

To analyze whether the impact of K-PAVE differed for Reading First and non-Reading First schools (research question 5), we used the same analytic approach described for subgroups of students. As for student subgroups, we built on the analytic model for testing impacts on students overall. To test impacts on subgroups of schools, we included an interaction of the intervention indicator with a school-level subgroup variable at level 3 of the hierarchical linear model (for example, T*RF, where RF = 1 for Reading First schools and 0 for non-Reading First schools) in equation (J8). The estimated parameter of such an interaction indicates whether there was a difference in impact of K-PAVE between Reading First and non-Reading First schools. To show this explicitly, we tested the interaction between the level 3 school characteristic and the level 3 treatment predictor by adding an interaction term to the level 3 equation:[57]

$$\text{(J19)}\ Y_{ijk} = \gamma_{000} + \gamma_{001}(T_k) + \gamma_{002}(RF_k) + \gamma_{003}(RF_k)(T_k)... + \upsilon_k$$

where γ_{001} indicates the average impact of the intervention for non-Reading First schools and γ_{003} indicates the differential in the average impact of K-PAVE between Reading First schools and non-Reading First schools. Therefore, the estimated average impact of the intervention for non-Reading First schools is the sum of γ_{001} and γ_{003}.

We conducted a *t*-test to test the null hypothesis that the differential in the average impact of K-PAVE for Reading First and non-Reading First schools was zero (that is, γ_{003} = 0) and used a .05-level criterion to reject the null hypothesis. A statistically significant estimate of γ_{003} would indicate that there is compelling evidence, at the 5% level, that the K-PAVE intervention had a differential impact on Reading First and non-Reading First schools. An estimate with a positive value would indicate that the impact of K-PAVE was larger for Reading First schools than for non-Reading First schools; a negative value would indicate that the impact of K-PAVE was smaller for Reading First schools than for non-Reading First schools. The magnitude of γ_{003} estimates the magnitude of the differential in the impact for Reading First schools compared with non-Reading First schools.

We calculated an effect size for the estimated differential in the impact (γ_{003}) by dividing the parameter estimate from the model by the standard deviation of the outcome variable in the control group for the full sample. Control group standard deviations were used, as recommended in Burghardt et al. (2009), because the intervention could affect the standard deviation in the intervention group.

[57] School-level covariates other than READING_FIRST are not shown in this model, in order to facilitate the illustration of the interaction between Reading First status and the treatment condition.

Two-level model used to estimate impacts on classroom instruction at the end of intervention year (kindergarten)

Controlling for teacher and school characteristics, we estimated the impact of the K-PAVE intervention on classroom instructional practices (for example, teacher lexical diversity, components of vocabulary and comprehension support) using a multilevel model, to account for the clustering of classrooms within schools. The model included a classroom level (level 1) and a school level (level 2). Because of the limited degrees of freedom at level 1 (because only two classrooms per school were sampled), teacher characteristics were controlled for at the school level. For each teacher characteristic, the average value for the school was calculated.

The multilevel model used to test impacts on classroom instruction was used to address two exploratory research questions (questions 6 and 8). 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 resulted 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.

The level 1, classroom-level, model is

(J20) $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$

where

- Y_{ij} is an outcome measure for the ith classroom in the jth school.
- β_{0j} is the mean value of the outcome Y for school j.
- ε_{ij} is the residual for the ith classroom in the jth school (the level 1 residuals were assumed to be normally distributed with mean 0 and variance σ^2).

The level 2, school-level, model is

$$\begin{aligned}
\beta_{0j} = {} & \gamma_{00} + \gamma_{01}(T_j) + \gamma_{02}(Tch\_AfrAm_{\mathrm{j}} - \overline{Tch\_AfrAm}) \\
& + \gamma_{03}(College_j - \overline{College}) + \gamma_{04}(GradDegree_j - \overline{GradDegree}) \\
& + \gamma_{05}(CertEC_{\mathrm{j}} - \overline{CertEC}) + \gamma_{06}(Certread_{\mathrm{j}} - \overline{Certread}) \\
& + \gamma_{07}(YrsTch_{\mathrm{j}} - \overline{YrsTch}) + \gamma_{08}(\mathrm{YrsTchKg}_{\mathrm{j}} - \overline{\mathrm{YrsTchKg}}) \qquad \text{(J21)} \\
& + \gamma_{09}(readingFirst_j - \overline{readingFirst}) + \gamma_{10}(MSStateInit_j - \overline{MSStateInit}) \\
& + \gamma_{11}(AchLvlIndex_j - \overline{AchLvlIndex}) + \gamma_{12}(PctAfrAm_j - \overline{PctAfrAm}) \\
& + \gamma_{13}(PctFreeLunch_j - \overline{PctFreeLunch}) + \gamma_{14}(SmallTown_j - \overline{SmallTown}) \\
& + \gamma_{15}(LrgeTown_j - \overline{LrgeTown}) + \gamma_{16}(Delta_j - \overline{Delta}) + \upsilon_{0j}
\end{aligned}$$

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_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.
- $\gamma_{02} - \gamma_{09}$ are regression coefficients indicating the effects of teacher characteristics on the outcome, averaged for school *j*.
- $\gamma_{10} - \gamma_{17}$ are regression coefficients indicating the effects of school characteristics on the outcome, for school *j*.
- $Tch\_AfrAm_j - \overline{Tch\_AfrAm}$ is the proportion of focal teachers in the *j*th school who are African American, centered at the grand mean.[58]
- $College_j - \overline{College}$ is the proportion of focal teachers in the *j*th school whose highest level of education is a bachelor's degree, centered at the grand mean.[59]
- $GradDegree_j - \overline{GradDegree}$ 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.
- $Certread_j - \overline{Certread}$ 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 years of teaching experience), 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, $\overline{MSStateInit}$.

[58] "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 were averaged for the two teachers from whom data were collected. Data on other kindergarten teachers in the school were not collected and thus were not included.

[59] The 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; *GradDegree* indicates that teachers' highest level of education is a graduate degree. The reference category is teachers with some graduate coursework but no graduate degree. In the sample, 38% of teachers had a bachelor's degree; 25% had some graduate coursework, and 36% had a graduate degree.

- $AchLvlIndex_j - \overline{AchLvlIndex}$ is the Achievement Level Index for the *j*th 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, $\overline{SmallTown}$.
- $LrgeTown_j$ 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, $\overline{LrgeTown}$. [60]
- $Delta_j$ is a dummy variable taking the value 1 if the ^{jth} school is in the Delta and 0 otherwise, centered at the grand mean, $\overline{Delta}$.
- υ_{oj} is the error term for the *j*th school (school-level error terms were assumed to be normally distributed with mean 0 and variance τ^2).

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

[60] 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.

APPENDIX K. SENSITIVITY ANALYSES

Sensitivity analyses were conducted to examine the robustness of impact estimates. These analyses were conducted for the confirmatory impact analysis reported in chapter 3, in which we examined the impacts of K-PAVE on students' expressive vocabulary, academic knowledge, and passage comprehension one year after the intervention ended, at the end of grade 1. Sensitivity analyses were also conducted for the exploratory analyses reported in appendix E, in which the impacts of K-PAVE on students' lexical diversity at the end of kindergarten, teachers' lexical diversity at the end of the intervention year, and the four components of the vocabulary and comprehension support composite were investigated. We did not conduct sensitivity analyses for subgroup differences in impacts at the end of kindergarten and the end of grade 1 (chapter 4) beyond those conducted for the overall main effects models. The results of sensitivity analyses presented in this appendix were compared with the main models presented in chapter 3 and appendix E.

CONFIRMATORY ANALYSIS OF IMPACTS ON STUDENTS ONE YEAR AFTER INTERVENTION

Students' expressive vocabulary at follow-up

For the Expressive Vocabulary Test-2nd Edition (EVT-2), the impact model reported in chapter 3 was a three-level model (school, classroom, and student; see appendix J for model specifications), with a treatment indicator included at the school level to estimate the average impact of K-PAVE on students' EVT-2 scores at follow-up. The impact estimate was adjusted for students' baseline EVT-2 test score, other student covariates, and school characteristics. Missing values for pretest and follow-up scores were imputed using single stochastic regression; missing values for school and student covariates were 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, outliers, and weighting to adjust for the loss of one treatment school. The sensitivity analysis models (described below) were compared with the final impact model in chapter 3.

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 were estimated as part of the sensitivity analyses and had the same structure and covariates as the final model.

Four models testing sensitivity to missing data imputation were estimated: one that does not impute any missing values for follow-up test, pretest, or covariates (that is, with a sample including only complete cases, or listwise deletion); one that does not impute missing values for the EVT-2 follow-up test and pretest (that is, with a sample including only students with both follow-up and pretest scores) but uses the dummy variable method for missing covariates; one that does not imputing missing values for covariates other than pretest scores (that is, with a sample including only students with no missing covariates) but imputes missing follow-up test values using single stochastic regression imputation; and one that simply drops incorrectly administered tests 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; the other excluded 12 students who were tested three weeks after all other student testing was completed.

One model testing sensitivity to crossovers was estimated. It excluded five students who transferred to another study school and crossed conditions and for whom posttest and follow-up data were available.

Two models were estimated to test sensitivity to movement to nonstudy schools during the kindergarten intervention year. One excluded 17 students (11 from the treatment group and 6 from the control group) who transferred to another school before pretest and were never tested. The other excluded 45 students (21 from the treatment group and 24 from the control group) who transferred to another school at any point during the intervention year, either before pretest or between pretest and posttest (this group included the 17 students who were never tested and the 28 students who were never tested after pretest).

One model testing sensitivity to scoring 0 on the EVT-2 pretest was estimated. It excluded students who had a 0 score. (No students had a score of 0 on the EVT-2 at follow-up.)

One model testing sensitivity to outliers was estimated. It excluded seven outliers. For the EVT-2 and other student outcomes, observations with studentized residuals that were more than 3.5 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 none of the sensitivity analysis models was the impact estimate statistically significant. Estimated magnitudes ranged from –0.35 to 0.52, estimated standard errors ranged from 0.68 to 0.76, and standardized effect sizes ranged from –0.03 to 0.05 (table K1). Findings were robust.

Students' academic knowledge and passage comprehension at follow-up

Sensitivity analyses were conducted to examine whether the findings regarding impacts on academic knowledge and passage comprehension were robust. Models were estimated to examine the sensitivity of the findings to the same factors examined for the EVT-2: covariate adjustment, missing data imputation, delayed baseline student testing, student crossovers, out-migration, outliers, and students with a score of 0 on the passage comprehension test.[61]

Findings regarding academic knowledge and passage comprehension were robust (tables K2 and K3). Impacts were not statistically significant in any models, using the $p < .025$ for statistical significance set by the Bonferroni adjustment. Even without the Bonferroni

[61] The sensitivity of findings to the presence of 0 scores on the academic knowledge test was not tested, because no students had raw scores of 0 on the academic knowledge pretest or follow-up test.

adjustment, no impacts were statistically significant; all *p* values were larger than .05 (.06–.27 for academic knowledge and .30–.96 for passage comprehension). For academic knowledge, unstandardized impact estimates ranged from 0.98 to 1.32, with standard errors of 0.82– 0.96 and standardized effect size estimates of 0.08–0.13. For passage comprehension, unstandardized impact estimates ranged from –1.71 to 0.69, with standard errors of 1.58–1.76 and standardized effect size estimates of–0.09 to 0.04.

Table K1. Estimated impact of K-PAVE on grade 1 students' Expressive Vocabulary Test-2 scores in final impact model and models fit for sensitivity analysis

Model	Impact estimate	Standard error	*t*–statistic	*p*–value	95% confidence interval		Effect size
Final model[a]	0.36	0.71	0.51	.61	–1.06	1.78	0.032
Covariate adjustment							
No covariates except treatment status and pretest	–0.35	0.76	–0.46	.65	–1.87	1.17	–0.03
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	0.18	0.76	0.24	.81	–1.34	1.70	0.02
Excluding cases with missing test scores	0.52	0.68	0.77	.44	–0.84	1.88	0.05
Excluding cases with missing covariates (other than pretest)	0.16	0.80	0.20	.84	–1.44	1.76	0.01
Excluding cases with incorrectly administered tests	0.35	0.71	0.49	.63	–1.07	1.77	0.03
Delayed baseline testing							
Excluding 27 students with baseline testing at least one week late	0.35	0.72	0.48	.63	–1.09	1.79	0.03
Excluding 12 students with baseline testing at least three weeks late	0.28	0.72	0.39	.70	–1.16	1.72	0.02
Crossovers and out–migrants							
Excluding five crossovers	0.39	0.71	0.55	.58	–1.03	1.82	0.04
Excluding transfers before baseline	0.39	0.71	0.55	.58	–1.03	1.81	0.03
Excluding transfers anytime during intervention year	0.47	0.71	0.66	.51	–0.95	1.88	0.04
Zero raw score							
Excluding students with zero raw score	0.31	0.70	0.44	.66	–1.10	1.71	0.03
Outliers							
Excluding seven outliers	0.14	0.68	0.20	.84	–1.22	1.50	0.01
Weighting schools							
Excluding school weights adjusting for school dropout	0.37	0.71	0.52	.60	–1.05	1.79	0.03

a. The final impact model is presented in chapter 3.

Table K2. Estimated impact of K-PAVE on grade 1 students' academic knowledge in final impact model and models fit for sensitivity analysis

Model	Impact estimate	Standard error	*t*–statistic	*p*–value	95% confidence interval		Effect size
Final model[a]	1.23	0.85	1.46	.14	–0.46	2.92	0.10
Covariate adjustment							
No covariates except treatment status and pretest	0.98	0.82	1.19	.23	–0.66	2.62	0.078
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	1.69	0.90	1.88	.06	–0.11	3.49	0.13
Excluding cases with missing test scores	1.33	0.79	1.68	.09	–0.25	2.91	0.11
Excluding cases with missing covariates (other than pretest)	1.58	0.96	1.65	.10	–0.34	3.50	0.13
Excluding cases with incorrectly administered tests	1.01	0.91	1.11	.27	–0.81	2.83	0.08
Delayed baseline testing							
Excluding students with baseline testing at least one week late	1.26	0.86	1.47	.14	–0.46	2.98	0.10
Excluding students with baseline testing at least three weeks late	1.24	0.85	1.46	.14	–0.45	2.94	0.10
Crossovers and out–migrants							
Excluding five crossovers	1.26	0.85	1.49	.14	–0.44	2.96	0.10
Excluding transfers before baseline	1.32	0.84	1.58	.12	–0.36	2.99	0.10
Excluding transfers anytime during intervention year	1.31	0.83	1.58	.11	–0.35	2.97	0.10
Outliers							
Excluding eight outliers	1.35	0.79	1.72	.09	–0.22	2.92	0.11
Weighting schools							
Excluding school weights adjusting for school dropout	1.24	0.84	1.47	.14	–0.45	2.92	0.10

a. The final impact model is presented in chapter 3.

Table K3. Estimated impact of K-PAVE on grade 1 students' passage comprehension in final impact model and models fit for sensitivity analysis

Model	Impact estimate	Standard error	*t*-statistic	*p*–value	95% confidence interval		Effect size
Final model[a]	0.50	1.69	0.29	.77	–2.89	3.88	0.025
Covariate adjustment							
No covariates except treatment status and pretest	0.09	1.66	0.05	.96	–3.24	3.42	0.004
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	–1.71	1.65	–1.04	.30	–5.01	1.59	–0.087
Excluding cases with missing test scores	–0.35	1.58	–0.22	.82	–3.51	2.81	–0.018
Excluding cases with missing covariates (other than pretest)	–0.71	1.76	–0.41	.68	–4.23	2.81	–0.036
Excluding cases with incorrectly administered tests	0.26	1.72	0.15	.88	–3.18	3.70	0.013
Delayed baseline testing							
Excluding students with baseline testing at least one week late	0.69	1.70	0.41	.69	–2.71	4.09	0.035
Excluding students with baseline testing at least three weeks late	0.47	1.70	0.28	.78	–2.93	3.87	0.024
Crossovers and out-migrants							
Excluding five crossovers	0.53	1.70	0.31	.76	–2.87	3.93	0.027
Excluding transfers missing pretest and posttest	0.43	1.70	0.25	.80	–2.97	3.82	0.022
Excluding transfers missing posttest	0.16	1.65	0.10	.92	–3.14	3.46	0.008
Zero raw score							
Excluding students with zero raw score	0.26	1.70	0.15	.88	–3.15	3.66	0.013
Outliers							
Excluding four outliers	0.30	1.68	0.18	.86	–3.05	3.66	0.015
Weighting schools							
Excluding school weights adjusting for school dropout	0.50	1.69	0.29	.77	–2.89	3.88	0.025

a. The final impact model is presented in chapter 3.

EXPLORATORY ANALYSES OF IMPACTS AT THE END OF THE INTERVENTION YEAR

Students' lexical diversity in kindergarten

We conducted a sensitivity analysis of the model testing the impact of K-PAVE on students' lexical diversity at the end of kindergarten (see appendix E). The impact model is a three-level model (school, classroom, and student; see appendix J for the model specifications), with an intervention indicator included at the school level to estimate the average impact of K-PAVE on students' lexical diversity score at posttest (that is, at the end of the kindergarten intervention year). The impact was adjusted for students' baseline lexical diversity score, other student covariates, and school characteristics. Missing values for pretest and posttest lexical diversity scores were imputed using single stochastic regression; missing values for school and student covariates were imputed using the dummy variable adjustment.

For student lexical diversity, 10 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, and weighting to adjust for the loss of one intervention school. The sensitivity analysis models (described below) were compared with the final impact model in appendix E.

One model testing sensitivity to covariate adjustment was estimated with no covariates other than the treatment indicator and students' lexical diversity pretest score. All other models were estimated as part of the sensitivity analyses and had the same structure and covariates as the final model.

Three models testing sensitivity to missing data imputation were estimated: one without imputing any missing values for follow-up test, pretest, or covariates (that is, with a sample including only complete cases, or listwise deletion); one without imputing missing values for the student lexical diversity scores at 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; and one 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 pretest and posttest lexical diversity scores using single stochastic regression imputation.

Two models testing sensitivity to delayed baseline testing were estimated. One excluded eight students tested more than one week after K-PAVE training was completed; the other excluded four students who were tested three weeks after all other student testing was completed.

One model testing sensitivity to crossovers was estimated. It excluded students who transferred to another study school and crossed conditions during the intervention year and for whom posttest data are available.

Two models were estimated to test sensitivity to movement to nonstudy schools during the kindergarten intervention year. One excluded students who transferred to another school before pretest and were never tested. The other excluded students who transferred to another school at any point during the intervention year, either before pretest or between pretest and posttest (this group included students who were never tested and students who were never tested after pretest).

One model testing sensitivity to outliers was estimated with outliers excluded. For the other student outcomes, observations with studentized residuals that were more than 3.5 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 none of the sensitivity analysis models was the impact estimate statistically significant. Estimated magnitudes ranged from –0.05 to 1.18, estimated standard errors ranged from 1.02 to 1.50, and standardized effect sizes ranged from –0.004 to 0.09 (table K4). Findings were found robust.

Teachers' lexical diversity at end of intervention year

We conducted a sensitivity analysis of the model testing the impact of K-PAVE on teachers' lexical diversity at the end of the intervention year (see appendix E). The impact model is a two-level model (classrooms, schools), with a treatment indicator included at the school level to estimate the average impact on the teacher lexical diversity outcome measure (see appendix J on model specifications). The impact estimate was adjusted for baseline teacher lexical diversity score, teacher characteristics, and school characteristics. Missing values for covariates other than the pretest were imputed using the dummy variable adjustment; missing lexical diversity scores were imputed using single stochastic regression.

The results of the sensitivity analyses were compared with those from the final models reported in appendix E that examine the impact on teacher lexical diversity at the end of the intervention year (table K5). Models were estimated to examine the sensitivity of the findings from the impact models to covariate adjustment, missing data imputation, outliers, and weighting to adjust for school dropout.

Table K4. Estimated impact of K-PAVE on kindergarten students' lexical diversity in final exploratory model and models fit for sensitivity analysis

Model	Impact estimate	Standard error	*t*-statistic	*p*–value	95% confidence interval		Effect size
Final model[a]	0.47	1.46	0.32	.75	–2.46	3.39	0.036
Covariate adjustment							
No covariates except treatment status and pretest	1.18	1.47	0.81	.42	–1.75	4.12	0.091
Missing data imputation							
Excluding cases with any missing data (listwise deletion)	0.63	1.02	0.62	.54	–1.41	2.67	0.049
Excluding cases with missing test scores	0.77	1.45	0.53	.60	–2.13	3.66	0.059
Excluding cases with missing covariates (other than pretest)	–0.05	1.50	-0.03	.97	–3.04	2.94	–0.004
Delayed baseline testing							
Excluding students with baseline testing at least one week late	0.28	1.47	0.19	.85	–2.67	3.22	0.021
Excluding students with baseline testing at least three weeks late	0.21	1.46	0.15	.88	–2.70	3.13	0.016
Crossovers and out-migrants							
Excluding crossovers	0.45	1.46	0.31	.76	–2.48	3.37	0.034
Excluding transfers before baseline	0.44	1.46	0.30	.76	–2.48	3.36	0.034
Excluding transfers anytime during intervention year	0.39	1.44	0.27	.79	–2.49	3.27	0.030
Outliers							
Excluding outliers	0.76	1.43	0.53	.60	–2.10	3.61	0.058
Weighting schools							
Excluding school weights adjusting for school dropout	0.47	1.46	0.32	.75	–2.46	3.40	0.034

a. The final exploratory impact model is presented in appendix E.

Table K5 presents the results of the sensitivity analysis, with impact estimates, standard errors, *t*-test results, 95% confidence intervals, and standardized effect sizes. The magnitude and standard error of the impact estimate remained consistent across all but one of the sensitivity models. In all models but one, the intervention impact was not statistically significant, with *p*-values of 0.08– 0.12. The estimated magnitudes were 3.6–4.1, estimated standard errors were 2.2–2.4, and standardized effect sizes were 0.30–0.34. The exception was a model without adjustment for teacher and school covariates (that is, controlling only for baseline score and intervention status), which suggested a larger standardized effect size of 0.48 (compared with 0.34 in the final model) and a statistically significant impact ($t = 2.74$, $p = .008$). This difference suggests that by not controlling for covariates, variation in lexical diversity associated with teacher and school characteristics may have been attributed to the K-PAVE intervention.

Components of classroom vocabulary and comprehension support at end of intervention

As part of the previous confirmatory analysis of K-PAVE impacts on the composite measure of vocabulary and comprehension support at the end of the intervention year, sensitivity analyses were conducted to examine the sensitivity of findings to covariates adjustment, missing data imputation, and weighting to adjust for the school that dropped out (see Goodson et al. 2010). Because these analyses had already been conducted, we did not repeat analyses examining sensitivity of findings to covariate adjustment, imputation of missing teacher and school covariates using the dummy variable adjustment, or weighting to adjust for the school that dropped out as part of the exploratory analysis of impacts on the four component variables used to create the composite.

The sensitivity of results to the imputation of missing baseline and posttest measures of the component variables was examined, as these missing values were imputed for the first time for the exploratory analyses. Three of the four variables used to create the vocabulary and comprehension support composite variable—comprehension support provided, higher-order questions asked, and vocabulary introduced during book read aloud—had some missing values. There were no missing values for the measure of vocabulary introduced during times other than the book read aloud. For each of the three variables with missing values, we compared one sensitivity model—a model with all covariates but without imputation of missing baseline and posttest values of the outcome—to the final model for each presented in appendix E.

The findings regarding the impact of K-PAVE on the components of the vocabulary and comprehension support remained consistent in the sensitivity analysis. In the sensitivity model for each outcome, the null hypothesis of a zero impact was rejected (table K6). The results were robust; the magnitude and standard errors of the estimated impact and the standardized effect size were nearly identical to those in the final models presented in appendix E.

Table K5. Sensitivity of estimated impact of K-PAVE on teachers' lexical diversity

Model	Estimate	Standard error	*t*-statistic	*p*–value	95% confidence interval		Effect size
Final model[a]	4.05	2.30	1.76	.08	–0.56	8.66	0.34
Covariate adjustment							
No covariates except treatment status and pretest	5.63	2.06	2.74	.01	1.51	9.74	0.48
Missing data imputation							
Missing covariates and posttest not imputed	3.94	2.27	1.73	.09	-0.60	8.49	0.33
Missing covariates not imputed	3.92	2.17	1.81	.08	-0.41	8.25	0.33
Missing posttest not imputed	3.76	2.39	1.57	0.12	-1.03	8.54	0.32
Outliers							
Excluding outliers	3.59	2.25	1.59	.12	–0.92	8.10	0.30
Weighting schools							
Excluding weights adjusting for school dropout	4.03	2.30	1.75	.09	-0.57	8.63	0.34

a. The final exploratory impact model is presented in appendix E.

Table K6. Sensitivity of estimated impact of K-PAVE on components of vocabulary and comprehension support to missing data imputation and outliers

Component	Estimate	Standard error	*t*-statistic	*p*–value	95% confidence interval		Effect size
Comprehension support during read aloud							
Final model[a]	10.91	3.33	3.27	.00	4.25	17.57	0.74
Missing posttest and baseline not imputed	10.89	3.31	3.29	.00	4.26	17.52	0.74
Excluding outliers	10.28	3.10	3.32	.00	4.08	16.48	0.70
Higher-order questions during read aloud							
Final model[a]	2.61	0.951	2.76	.01	0.72	4.51	0.80
Missing posttest and baseline not imputed	2.62	0.95	2.75	.01	0.72	4.52	0.79
Excluding outliers	2.18	0.73	2.98	.00	0.72	3.64	0.66
Vocabulary during read aloud							
Final model[a]	2.09	0.92	2.26	.03	0.24	3.94	0.50
Missing posttest and baseline not imputed	2.09	0.91	2.30	.03	0.27	3.90	0.50
Excluding outliers	2.26	0.90	2.52	.01	0.47	4.06	0.54
Vocabulary during times other than read aloud[b]							
Final model [a]	0.54	0.33	1.63	0.11	-0.12	1.21	.38
Excluding outliers	0.61	0.31	1.95	.06	-0.02	1.23	.43

a. The final exploratory impact model is presented in appendix E.

b. There were no missing values for one of the four components of the vocabulary and comprehension support composite—vocabulary introduced during times of the day other than the read aloud; therefore, sensitivity of results to missing data imputation was not required for this outcome.

APPENDIX L. IMPUTATION OF MISSING DATA

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 were imputed separately for treatment and control groups.

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/ethnicity.
- Eligibility for free or reduced-price meals.
- Special education status (having an Individualized Education Plan).

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

Missing data on school covariates

Data were collected for the following school covariates:

- Reading initiative in place before K-PAVE (Reading First, a Mississippi state initiative, or other).
- Achievement Level Index.[62]
- Percentage of African American students.
- Percentage of students eligible for free or reduced-price meals.
- Locale type (rural, small town, large town/fringe of city).
- Location (within or contiguous with the Delta).

There were no missing data in either the treatment or control group on any of the school covariates except the Achievement Level Index (table L1).

[62] 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 (the Mississippi Curriculum Test), which is administered to all students in grades 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 scoring 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."

Table L1. Missing data on school covariates

	Treatment schools (*n* = 30)		Control schools (*n* = 34)	
Covariate	**Number**	**Percent**	**Number**	**Percent**
Reading initiative	0	0	0	0
Achievement Level Index	4	13.3	3	8.8
Percentage of African American students	0	0	0	0
Percentage of students eligible for free or reduced-price meals	0	0	0	0
Locale type	0	0	0	0
Location	0	0	0	0

Missing data on teacher covariates

Data were collected on the following teacher covariates:

- Race/ethnicity.
- 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.

Data were missing on at least one covariate for 3.1% of teachers (4 of 128).

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, all missing values were set to a constant value of 0. In addition, the analysis included an indicator variable identifying observations for which the value of the covariate was missing. The indicator variable for a given covariate was set to 1 where the covariate was missing and to 0 where the covariate was not missing. Although some research (for example, Jones 1996) shows that this method generally produces biased estimates, a recent National Center for Education Evaluation and Regional Assistance Technical Methods report finds that it performed well in simulations that mirrored the randomized study design and analysis plan of this evaluation (Puma et al. 2009).

The 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 the 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, posttest, and follow-up data

The percentage of students missing pretest assessment data was 4%–6% in the intervention group and 3%–6% in the control group. The percentage of students missing posttest assessment data was 4%–6% in the intervention group and 5%–8% in the control group. At follow-up, 12% of students in the intervention group and 13% in the control group were missing assessments (table L2).

Table L2. Missing data on student pretest, posttest, and follow-up assessments, for intervention and control groups

	Baseline		Kindergarten posttest		Grade 1 follow-up	
Group/component	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Intervention group *(n = 596 students)*						
Expressive vocabulary	22	3.7	25	4.2	74	12.4
Academic knowledge	22	3.7	26	4.4	74	12.4
Listening comprehension	36	6.0	25	4.2	a	a
Passage comprehension	22	3.7	a	a	74	12.4
Student lexical diversity	14	5.7	15	6.1	a	a
Control group *(n = 700 students)*						
Expressive vocabulary	24	3.4	38	5.4	90	12.9
Academic knowledge	25	3.6	38	5.4	90	12.9
Listening comprehension	25	3.6	39	5.6	a	a
Passage comprehension	24	3.4	a	a	91	13.0
Student lexical diversity	16	5.7	23	8.2	a	a

a. Not assessed.

Five classroom instruction measures—teachers' lexical diversity and the four component variables used to create the vocabulary and comprehension support composite (comprehension support during book read aloud, higher-order questions during read aloud, vocabulary during read aloud, and vocabulary during times other than the read aloud)—were examined in the exploratory analyses. Classroom instruction measures were collected at baseline and kindergarten posttest only. Table L3 shows rates of missing data for baseline and posttest measures of classroom instruction, for intervention and control groups.

Table L3. Percentage of missing data on classroom instruction measures at pretest and posttest

Test	Intervention group (*n* = 60 classrooms)	Control group (*n* = 68 classrooms)
Pretest	0–12	0–2
Posttest	0–3	0–6

Rates of missing data were higher for the lexical diversity measure at baseline than at posttest because some teachers were uncomfortable being audio recorded. Rates of missing teacher lexical diversity at baseline were higher in the intervention group than the control group (*p* = .03). At posttest, a higher percentage of teachers were recorded; rates of missing data did not differ for intervention and control groups (*p* = .68).

For student assessments and classroom instruction measures, missing pretest, posttest, or follow-up scores were imputed using single stochastic regression (see Puma et al. 2009 for a description).[63] We adjusted a multiple regression model for the multilevel structure of the data to estimate predicted values for each pretest, posttest, or follow-up measure with missing values. Predictors in each imputation model included all other available information collected (including pretest scores, posttest scores, follow-up scores, and covariates). For each pretest, posttest, and follow-up 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 achieve the same variation in imputed values as in observed values.

The imputation of missing student assessments and classroom instruction measures was done in two stages. For the kindergarten confirmatory impact analysis, missing pretest and posttest scores were imputed using single stochastic regression. Imputed values for missing pretest and posttest scores were used in the kindergarten confirmatory impact analysis reported in Goodson et al. 2010. A second round of missing data imputation was conducted for the follow-up analysis, in which missing student assessments and classroom instruction measures that were not previously imputed for the kindergarten confirmatory analysis were imputed for the confirmatory analysis of impacts at follow-up and the exploratory analysis of kindergarten and grade 1 impacts. In the second stage, missing data for the following variables were imputed:

- Follow-up student assessment measures: expressive vocabulary, academic knowledge, and passage comprehension.
- Baseline passage comprehension.
- Baseline and posttest student lexical diversity.

[63] 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.

- Baseline and posttest teacher lexical diversity.
- Posttest scores for three of the vocabulary and comprehension support variables: comprehension support provided, higher-order questions posed, and vocabulary introduced during read aloud.

In both stages, all known variables were included as predictors of variables with missing values. When imputing pretest scores in the first stage, we included posttest scores 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 missing Expressive Vocabulary Test (EVT) pretest scores, all covariates were used to obtain predicted values, and Expressive Vocabulary Test 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 is in turn used to predict the posttest outcome in the impact model— experts recommend this approach (Little and Rubin 2002; Moons, Donders, Stijnen, and Harrell 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, we included pretest and posttest scores for other student outcomes. 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 posttest measures. Following the recommendations of Puma et al. (2009), the imputation models included any measured variables that may be associated with missing data.

In the second stage of imputing missing data, pretest and posttest scores that were imputed in the first stage were included as predictors of missing scores imputed in the second stage.

Equation L1 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, including student lexical diversity. A series of 29 treatment school dummy variables was included in the model to adjust for the nesting of students in 30 treatment schools.[64] 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.

[64] Rather than estimating a multilevel model for imputing missing values, we used a series of dummy variables to adjust for the nested structure of the data.

(L1)

$$\begin{aligned}
EVT\_pretest = {} & \beta_0 + \beta_1 EVT\_posttest \\
& + \beta_2 AcadKnow\_pretest + \beta_3 AcadKnow\_posttest \\
& + \beta_4 ListeningComp\_pretest + \beta_5 ListeningComp\_posttest \\
& + \beta_6 Stud\_PosttestAge + \beta_7 Stud\_Female \\
& + \beta_8 Stud\_AfrAm + \beta_9 Stud\_EligFreeLunch \\
& + \beta_{10} Stud\_IEP + \beta_{11} Sch\_readingFirst + \beta_{12} Sch\_MSStateInit \\
& + \beta_{13} Sch\_AchLvlIndex + \beta_{14} Sch\_PctAfrAm \\
& + \beta_{15} Sch\_PctEligFreeLunch + \beta_{16} SmTown + \beta_{17} LgTown \\
& + \beta_{18} Delta + \alpha_1 TreatmentSch1 + \ldots + \alpha_{29} TreatmentSch29 + \varepsilon.
\end{aligned}$$

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 L2 shows the model used to predict missing student assessment posttest scores, using the Expressive Vocabulary Test posttest for the control group as an example. Thirty-three 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.

(L2)

$$\begin{aligned}
EVT\_posttest = {} & \beta_0 + \beta_1 EVT\_pretest \\
& + \beta_2 AcadKnow\_pretest + \beta_3 AcadKnow\_posttest \\
& + \beta_4 ListeningComp\_pretest + \beta_5 ListeningComp\_posttest \\
& + \beta_6 Stud\_PosttestAge + \beta_7 Stud\_Female \\
& + \beta_8 Stud\_AfrAm + \beta_9 Stud\_EligFreeLunch \\
& + \beta_{10} Stud\_IEP + \beta_{11} Sch\_readingFirst + \beta_{12} Sch\_MSStateInit \\
& + \beta_{13} Sch\_AchLvlIndex + \beta_{14} Sch\_PctAfrAm \\
& + \beta_{15} Sch\_PctEligLunch + \beta_{16} SmTown + \beta_{17} LgTown \\
& + \beta_{18} Delta + \alpha_1 ControlSch1 + \ldots + \alpha_{33} ControlSch33 + \varepsilon.
\end{aligned}$$

Equation L3 shows the model used to predict missing follow-up student assessments, using expressive vocabulary at posttest for control group schools as an example. There were 33 control school dummy variables 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. In this second stage of missing data imputation, missing pretest and posttest scores that were imputed during the first stage were included as predictors in the imputation models.

(L3)

$$
\begin{aligned}
EVT\_followup = {} & \beta_0 + \beta_1 EVT\_pretest\_imp + \beta_2 EVT\_posttest\_imp \\
& + \beta_3 AcadKnow\_pretest\_imp \\
& + \beta_4 AcadKnow\_posttest\_imp \\
& + \beta_5 AcadKnow\_followup \\
& + \beta_6 ListeningComp\_pretest\_imp \\
& + \beta_7 ListeningComp\_posttest\_imp \\
& + \beta_8 ListeningComp\_followup \\
& + \beta_9 PassageComp\_pretest + \beta_{10} PassageComp\_posttest \\
& + \beta_{11} Stud\_FollowupAge + \beta_{12} Stud\_Female \\
& + \beta_{13} Stud\_AfrAm + \beta_{14} Stud\_EligFreeLunch \\
& + \beta_{15} Stud\_IEP + \beta_{16} Sch\_readingFirst + \beta_{17} Sch\_MSStateInit \\
& + \beta_{18} Sch\_AchLvlIndex + \beta_{19} Sch\_PctAfrAm \\
& + \beta_{20} Sch\_PctEligFreeLunch + \beta_{21} SmTown + \beta_{22} LgTown \\
& + \beta_{23} Delta + \alpha_1 ControlSch1 + ... + \alpha_{33} ControlSch33 + \varepsilon.
\end{aligned}
$$

The same approach was used to impute missing values for the classroom instruction pretest and posttest measures (equations L4 and L5) examined in the exploratory analyses—teacher lexical diversity and components of the vocabulary and comprehension support composite (three of which had missing values at posttest). Predictors in the model were school covariates, teacher covariates, classroom instruction baseline and posttest measures, and a series of 33 control school dummy variables to adjust for the nesting of classrooms 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.

(L4)

$$
\begin{aligned}
TchrLexicalD\_pretest = \beta_0 &+ \beta_1 TchrLexicalD\_posttest \\
&+ \beta_2 VocCompSup\_pretest \\
&+ \beta_3 VocCompSup\_posttest\_imp \\
&+ \beta_4 InstrSup\_pretest + \beta_5 InstrSup\_posttest \\
&+ \beta_6 EmotSup\_pretest + \beta_7 EmotSup\_posttest \\
&+ \beta_8 OthLit\_pretest + \beta_9 OthLit\_posttest \\
&+ \beta_{10} Tch\_Female + \beta_{11} Tch\_AfrAm + \beta_{12} College \\
&+ \beta_{13} GradDegree + \beta_{14} CertEC + \beta_{15} Certread \\
&+ \beta_{16} YrsTch + \beta_{17} YrsTchKg + \beta_{18} Sch\_readingFirst \\
&+ \beta_{19} Sch\_MSStateInit + \beta_{20} Sch\_AchLvlIndex \\
&+ \beta_{21} Sch\_PctAfrAm \\
&+ \beta_{22} Sch\_PctEligFreeLunch \\
&+ \beta_{23} SmTown + \beta_{24} LgTown + \beta_{25} Delta \\
&+ \alpha_1 ControlSch1 + ... + \alpha_{33} ControlSch33 + \varepsilon.
\end{aligned}
$$

(L5)

$$
\begin{aligned}
TchrLexicalD\_posttest = \beta_0 &+ \beta_1 TchrLexicalD\_pretest \\
&+ \beta_2 VocCompSup\_pretest \\
&+ \beta_3 VocCompSup\_posttest\_imp \\
&+ \beta_4 InstrSup\_pretest + \beta_5 InstrSup\_posttest \\
&+ \beta_6 EmotSup\_pretest + \beta_7 EmotSup\_posttest \\
&+ \beta_8 OthLit\_pretest + \beta_9 OthLit\_posttest \\
&+ \beta_{10} Tch\_Female + \beta_{11} Tch\_AfrAm + \beta_{12} College \\
&+ \beta_{13} GradDegree + \beta_{14} CertEC + \beta_{15} Certread \\
&+ \beta_{16} YrsTch + \beta_{17} YrsTchKg + \beta_{18} Sch\_readingFirst \\
&+ \beta_{19} Sch\_MSStateInit + \beta_{20} Sch\_AchLvlIndex \\
&+ \beta_{21} Sch\_PctAfrAm \\
&+ \beta_{22} Sch\_PctEligFreeLunch \\
&+ \beta_{23} SmTown + \beta_{24} LgTown + \beta_{25} Delta \\
&+ \alpha_1 ControlSch1 + ... + \alpha_{33} ControlSch33 + \varepsilon.
\end{aligned}
$$

Once all missing values were imputed using the single stochastic regression approach, the confirmatory and exploratory student and classroom impact models were estimated using imputed values. As a form of sensitivity analysis, impact models were also estimated without imputed values for missing data (that is, only with nonmissing cases). Sensitivity analyses (reported in appendix K) indicated that the magnitude and standard errors of impact estimates were similar regardless of whether missing pretest, posttest data, or covariate data were imputed. The imputation of missing data did not affect whether impact estimates were statistically significant.

APPENDIX M. 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 any time 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 Process

I'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 OBSERVATIONS

When 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 her to wear the recorder around her neck and that the clip secures the recorder so that it does not swing. You can work with the teacher to help her 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 N. DATA QUALITY ASSURANCE PROCEDURES

TRAINING STUDENT ASSESSMENT DATA COLLECTORS

Student assessment data collectors were responsible for administering tests to students. The SERVE Center at the University of North Carolina at Greensboro recruited assessors from local universities and community colleges in Mississippi and through contacts at the Mississippi Department of Education. Data collectors included college students with a background in education, retired teachers, school counselors, and school administrators. Data collectors were independent of the intervention implementation and unaware of the intervention status of a school. As a result, data collection procedures could be maintained as identical in intervention and control schools. (Student assessment procedures are described in appendix M.)

Data collectors 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 numerous opportunities to practice mock test administrations.[65]

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).

Twenty-one of 34 trainees (62%) were certified to collect baseline data. For posttest data collection, both returning and new data collectors were trained and certified. All 13 returning assessors and 6 of 9 new trainees were certified to collect data. For follow-up data collection, 16 returning assessors attended refresher training; all were recertified to collect data.

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. No measure of interrater reliability was collected during these visits. Within the first week of data collection, trainers held conference calls with data collectors to discuss questions and challenges.

[65] For all child assessments—the Expressive Vocabulary Test-2, the Woodcock-Johnson III/NU Academic Knowledge and Passage Comprehension tests, and the 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 the 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.

Quality assurance protocol for student assessments

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 level 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. 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; 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 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, raw scores were converted to *W*-scores, which are Item Response Theory–based scale scores, using the Woodcock-Johnson Compuscore Program, which is included in the WJIII/NU test kit (Woodcock et al. 2007). 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, we compared the *W*-scores with raw scores from the original data file; any inconsistencies were flagged, double-checked, and rectified, if necessary. Classroom instruction variables, which were created by summing and averaging observer ratings or tallies, were 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 L), 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.[66] Hard-copy score forms and assessor notes were examined to confirm that the 0 raw scores were accurate for each of these students. Because there were no notes from assessors indicating that the student refused to

[66] Raw scores of 0 for listening comprehension were not unusual.

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 K).

APPENDIX O. SCHOOL, TEACHER, AND STUDENT COVARIATES

Student-level covariates used in the analysis are defined in table O1.[67]

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, Passage comprehension)
Female (*Female*)	1 = student is female 0 = student is male
Special education status (student has Individualized Education Plan) (*StudentIEP*)	1 = student has Individualized Education Plan 0 = student does not have Individualized Education Plan
Eligibility for free or reduced-price meals (*FreeLunch*)	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

[67] We planned to include two other covariates in the analysis: whether students had been retained in kindergarten and whether students had attended preschool before entering kindergarten. Inadequate data quality precluded their use. Schools reported inconsistent information on retention. Some schools reported whether students had been retained 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 the subsequent year). Because comparable data were not available on all students, kindergarten retention status could not be included in the analysis. Schools also reported inconsistent information regarding whether students attended preschool before kindergarten. Some schools reported information about the type of institution children attended before kindergarten (prekindergarten, Head Start, day care, family child care, public/private). 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 child care in their definition of attending preschool; others may have included only prekindergarten and Head Start, excluding even programs in day care 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 (table O2) were averaged for each school and treated as covariates in the school-level model.

Table O2. Teacher characteristics used as school-level covariates

Covariate	How measured at teacher level	Averaged at school level
African American (*Tch_AfrAm*)	1 = teacher is African American 0 = teacher is not African American (for this sample, teacher is White)	1.0 = 100% of teachers are African American *(both teachers are African American)* 0.5 = 50% of teachers are African American *(one teacher is African American; one teacher is White)* 0.0 = 0% of teachers are African American *(i.e., both teachers are White)*
Level of education	Teacher's highest level of education completed, represented by a series of two dummy variables. Reference category is "some graduate courses". • Bachelor's degree (*College*) 1 = highest level of education is a bachelor's degree 0 = highest level of education is not a bachelor's degree • Graduate degree (*GradDegree*) 1 = highest level of education is a graduate degree 0 = highest level of education is not a graduate degree	Average of education level for two study teachers, represented by two variables. Reference category is "100% of teachers have some graduate courses". • Bachelor's degree (*College*) 1.0 = for 100% of teachers, bachelor's degree is highest education 0.5 = for 50% of teachers, bachelor's degree is highest education 0.0 = for 0% of teachers, bachelor's degree is highest education • Graduate degree (*GradDegree*) 1.0 = for 100% of teachers, graduate degree is highest education 0.5 = for 50% of teachers, graduate degree is highest education 0.0 = for 0% of teachers, graduate degree is highest education
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	1.0 = 100% of teachers have early childhood certification 0.5 = 50% of teachers have early childhood certification 0.0 = 0% of teachers have early childhood certification
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	1.0 = 100% of teachers have reading certification 0.5 = 50% of teachers have reading certification 0.0 = 0% of teachers have reading certification
Years teaching (*YrsTch*)	Number of years teacher has been teaching children (values are on a continuous scale)	Average number of years that the two study teachers have been teaching (continuous scale)
Years teaching kindergarten (*YrsTchKg*)	Number of years teacher has been teaching kindergarten (are on a continuous scale)	Average number of years that the two study teachers have been teaching kindergarten (continuous scale)

School-level covariates used in the analysis are defined in table O3.

Table O3. School covariates

Covariate	How measured
Treatment status (*T*)	1 = school was randomly assigned to receive K-PAVE treatment 0 = school was randomly assigned to control group
Reading initiatives	Represented by series of two dummy variables. Reference category is schools that have neither Reading First nor a Mississippi state reading initiative. Reading First (*ReadingFirst*): 1 = school participates in Reading First program 0 = school does not have Reading First program State reading initiative (*MSStateInit*) 1 = school has state reading initiative (for example, Barksdale, Reading Sufficiency) 0 = school does not have state reading initiative *Note: In subgroup analysis of differences in impacts on student outcomes for Reading First and non-Reading First schools, only one dummy variable –* ReadingFirst *– is included in the analysis model.*
Achievement Level Index (*AchLvlIndex*)	Measure of school-level achievement based on average scores on Mississippi Curriculum Test, given annually to all students in grades 3 and higher. Scores are on 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 type	Represented by series of two dummy variables. Reference category is *rural*. Small town (*SmallTown*) 1 = school is in a small town 0 = school is not in a small town Large town or fringe of city (*LargeTown*) 1 = school is in a large town or on fringe of a city 0 = school is not in a large town or on fringe of a city
Location (*Delta*)	1 = school is in Delta region 0 = school is in a county contiguous with Delta region

APPENDIX P. UNADJUSTED SAMPLE MEANS AND STANDARD DEVIATIONS FOR OUTCOME MEASURES

Table P1 displays unadjusted sample means and standard deviations for grade 1 student outcomes, lexical diversity in kindergarten, and teacher outcome measures examined in follow-up exploratory analyses, for students in both intervention and control schools.

Table P1. Unadjusted sample intervention and control group means for selected outcome measures

	Intervention group (*n* = 596 students)		**Control group (*n* = 700 students)**	
Measure	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Grade 1 student outcomes				
Expressive vocabulary, standard score	91.5	10.5	91.8	11.2
Academic knowledge, *W*-score	466.5	11.8	465.6	12.5
Passage comprehension, *W*-score	451.7	21.6	451.9	19.9
Kindergarten student outcome				
Lexical diversity, *D*	36.9	11.6	36.1	13.0
Kindergarten teacher outcomes				
Comprehension support (during read aloud), frequency	28.8	16.5	17.0	14.7
Higher-order questions (during read aloud), frequency	5.9	5.2	2.9	3.3
Vocabulary words (during read aloud), frequency	6.4	4.1	4.3	4.2
Vocabulary words (per 20-minute observation cycles during times other than reading), frequency	2.6	1.6	1.8	1.4
Teacher lexical diversity, *D*	84.9	12.2	80.1	11.9

Appendix Q. Classroom observation measures used to create vocabulary and comprehension support outcome variables

This appendix discusses two classroom observation measures used to create the vocabulary and comprehension support outcome variables examined in the exploratory analysis reported in appendix E. The classroom observation measures are the Read Aloud Profile–Kindergarten (RAP-K) and the Vocabulary Record. The appendix also describes the vocabulary and comprehension support composite, created from two RAP-K variables and two Vocabulary Record variables and analyzed in the kindergarten study (Goodson et al. 2010).

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 et al. 2004). The original instrument was adapted to use during book readings to focus primarily on teachers' comprehension support statements and questions, open-ended questions, and 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 to 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, and 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 promote children's comprehension and higher-order thinking abilities (see review of shared reading interventions in chapter 4 of National Early Literacy Panel 2008).[68]

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. It describes two aspects of the reader's strategies during the read aloud: the use of comprehension supports before, during, and after the text reading and the use of higher-order, cognitively challenging questions (figure Q1).

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

[68] 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 National Early Literacy Panel 2008). 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.

Assessment Scoring System (CLASS; Pianta, La Paro, and Hamre 2008) until the read aloud ended. Throughout the entire book reading, from the time the teacher announced she was going to read a book until any postreading discussion of the book was concluded, the observer documented 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 questions asking students to analyze, explain, predict, imagine, make inferences, or generate hypotheses. If the teacher did not read aloud to a group of students during the classroom observation, all behaviors were coded as occurring zero times; they were not coded as missing. If the teacher read aloud to a group of students more than once during the observation, the observer continued coding the CLASS for all additional instances of reading aloud; the RAP-K was coded only the first time the teacher read aloud.

VOCABULARY RECORD

The Vocabulary Record was adapted from the vocabulary component of the Instructional Practice in Reading Inventory (Smith et al. 2005). 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; 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 to the Vocabulary Record involved multiple iterations, involving both piloting early versions in kindergarten classrooms and coding 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 was completed during each Classroom Assessment Scoring System (CLASS) (Pianta, La Paro, and Hamre 2008) observation cycle and during the read aloud coded with the RAP-K.[69] The observer recorded all words defined by the teacher or assistant teacher and all words the teacher or assistant teacher asked a student to define (figure Q2). 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:

[69] The CLASS is a time-sampling observation tool used to rate the quality of interactions and instruction in kindergarten to grade 3 classrooms. It 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 of 10 dimensions of instruction). The Vocabulary Record was coded during the 20-minute observation portion of each CLASS cycle that was observed. On average, four cycles were coded in each classroom at baseline and at kindergarten posttest.

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

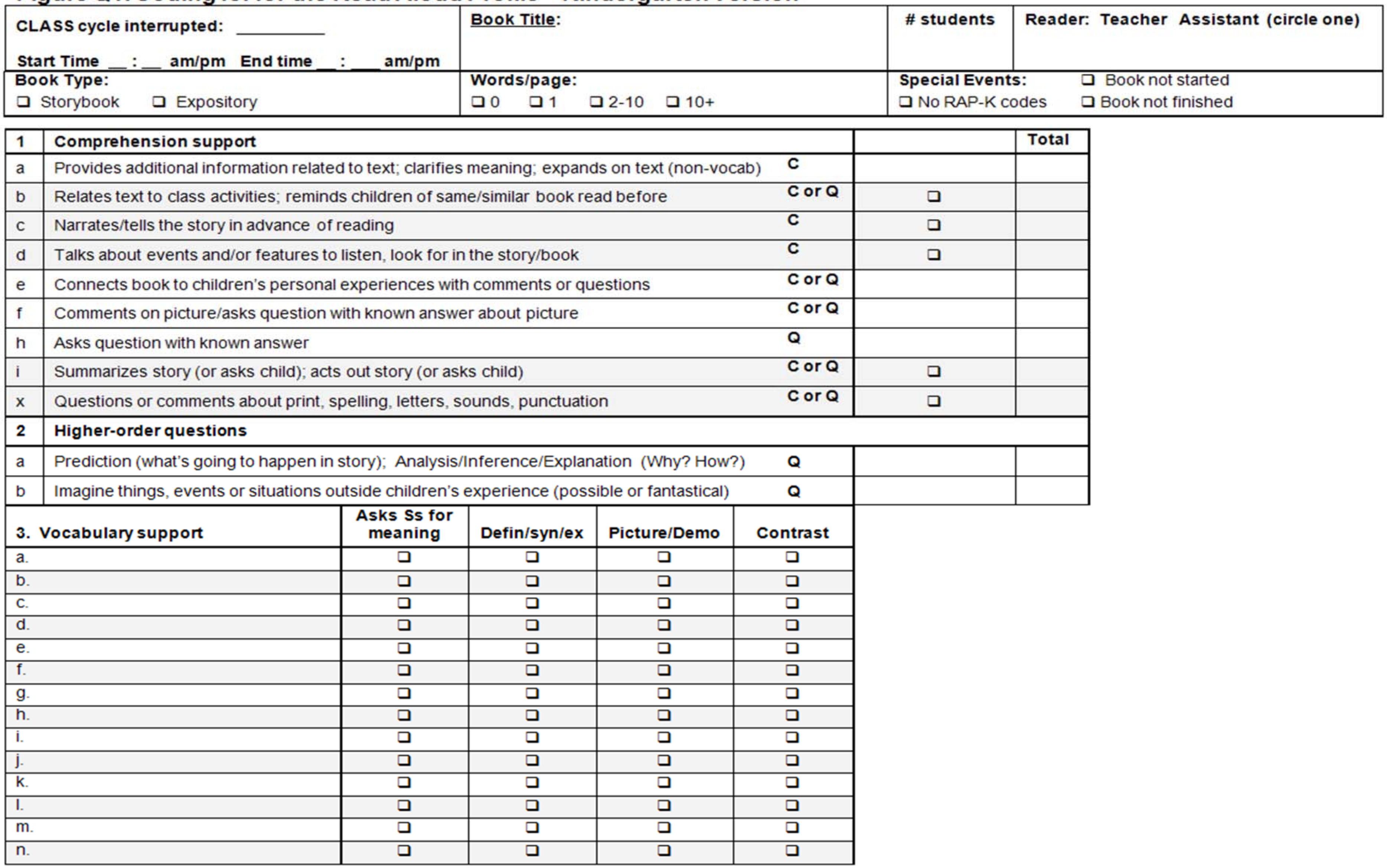
For this study, this instrument yielded two variables. The total number of words documented during the read aloud was tallied for a measure of the number of words introduced during the book reading. The total number of words documented during each CLASS cycle was tallied and averaged (that is, divided 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 (the 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 (the number of comprehension supports provided during reading and the number of higher-order questions posed 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 level of vocabulary and comprehension support was provided in the classrooms and a score of 1 indicates that the level of vocabulary and comprehension support is one standard deviation higher than average. A positive score indicates that a higher than average level of vocabulary and comprehension support is provided; a negative score indicates that a lower than average level of comprehension support is provided.

Cronbach’s alpha for the vocabulary and comprehension support composite was 0.62. Each variable created from the read aloud—comprehension support, higher-order questions, and number of vocabulary words—had a correlation with the composite of 0.45–0.55. Removing any of the variables from the composite reduced the internal consistency of the composite (that is, Cronbach’s alpha became 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 0.16. Removing this variable from the composite increased the internal consistency, raising Cronbach’s alpha to 0.71). However, all four variables were retained in the composite for analysis. The composite was determined based on the theorized relationship among the four variables as part of single domain before any examination of the relationships observed in the sample data. The aim was to avoid allowing an empirically defined composite to be overly influenced by sampling variation.

Figure Q1. Coding for for the Read Aloud Profile – Kindergarten Version



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.	□	□	□	□

Figure Q2. Vocabulary Record coding form

CLASS cycle # 1 Vocabulary support		Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast		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 Vocabulary support		Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast		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 **Vocabulary support**		**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**		**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 **Vocabulary support**		**Asks for meaning**	**Definition, example, or synonym**	**Picture or demonstration**	**Contrast**		**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

	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast		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

	Vocabulary support	Asks for meaning	Definition, example, or synonym	Picture or demonstration	Contrast		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 R. 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 (for example 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)
O 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)

References

Adams, M. J., Foorman, B. R., Lundberg, I., & Beeler, T. (1998). *Phonemic awareness in young children*: *A classroom curriculum*. Baltimore, MD: 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*, 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*, 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*, 498–520.

Bishaw, A., & Iceland, J. (2003). *Poverty*: *1999.* Census 2000 Brief. Washington, DC: U.S. Census Bureau.

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*, 465–473.

Burghardt, J., Deke, J., Kisker, E., Puma, M., & Schochet, P. (2009). *Regional educational laboratory rigorous applied research studies*: *Frequently asked analysis questions.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, and 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., & 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*, 1–18.

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*, 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.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*, 220–242.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2*, 1–45.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24*, 174–186.

Fishman, G. S., & Moore, L. R. (1982). A statistical evaluation of multiplicative congruential generators with modulus (2^{31} – 1). *Journal of the American Statistical Association, 77*(1), 29–136.

Forder, P. M., Gebski, V. J., & Keech, A. C. (2005). Allocation concealment and blinding: When ignorance is bliss. *Medical Journal of Australia, 182*(2), 87–89.

Glazerman, S., Levy, D. M., & Meyers, D. (2003). Nonexperiemental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science, 589*(1), 63–93.

Goffman, L., & Leonard, J. (2000). Growth of language skills in preschool children with specific language impairment: Implications for assessment and intervention. *American Journal of Speech-Language Pathology, 9*(2), 151–161.

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.

Goodson, B., Wolf, A., Bell, S., Turner, H., and Finney, P. B. (2010).*The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB).* (NCEE 2010-4014). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Hamilton, C. E., & Schwanenflugel, P. J. (2011). *PAVEd for success*: *A program for the development of vocabulary and oral language.* 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.

Houghton Mifflin Reading. (2008). Orlando, FL: Houghton Mifflin Harcourt School Publishers.

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 (KTEA-II)* (2nd ed.). Shoreview, MN: AGS Publishing.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John W. Wiley and Sons.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80*, 300–335.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–337.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary*: *Description, acquisition and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Mississippi Department of Education. (n.d.). *Mississippi curriculum test [proficiency levels].* Retrieved November 16, 2010, from http://www.mde.k12.ms.us/acad/osa/gltp.html

Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent-child book readings: A meta-analysis. *Early Education & Development*, *19*(1), 7–26.

Mol, S. E., Bus, A. G., & De Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research, 79*, 979–1007.

Mood, A., Graybill, F., & Boes, D. (1950). *Introduction to the theory of statistics.* New York: McGraw-Hill Companies.

Moons, K., Donders, R., Stijnen, T., & Harrell, F. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology, 59*, 1092–1101.

National Early Literacy Panel. (2008). *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*, 23–33.

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: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Quality Counts 2008. Special supplement. Mississippi state highlights. (2008). Bethesda, MD: Editorial Projects in Education Research Center.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*: *Applications and data analysis methods* (2nd Ed.). Newbury Park, CA: Sage.

Ready, D. D., LoGerfo, L. F., Lee, V. E., & Burkam, D.T. (2005). Explaining girls' advantage in kindergarten literacy learning: Do classroom behaviors make a difference? *Elementary School Journal, 106*, 21–38.

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*, 54–64.

Schochet, P. Z. (2008a). *Guidelines for multiple testing in impact evaluations. Technical methods report* (NCEE 2008-4018). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved November 16, 2010, from http://ies.ed.gov/ncee/pdf/20084018.pdf

Schochet, P. Z. (2008b). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*, 62–87.

Schwanenflugel, P. J., Hamilton, C. E., Neuharth-Pritchett, S., Restrepo, M. A., Bradley, B. A., & Webb, M. Y. (2010). PAVEd for Success: An evaluation of a comprehensive preliteracy program for four-year-old children. *Journal of Literacy Research*, *42*(3), 227–275.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals, (CELF-4)* (4th ed.). 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*, 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*, 218–229.

Silverman, S., & Ratner, N. B. (2002). Measuring lexical diversity in children who stutter: Application of vocd. *Journal of Fluency Disorders, 27*, 289–304.

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). New York: Russell Sage Foundation.

Smith, W. C., Dwyer, M. C., Dixon, Q., Schimmenti, J., Boulay, B., Khalil, B., et al. (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*, 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*, 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 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*, 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.

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*, 281–290.

Whitehurst, G.J., & Lonigan, C.J. (2001). Emergent Literacy: Development from Prereaders to Readers. In S. B Neuman & D. K. Dickensen (Eds.), *Handbook of Early Literacy Research* (pp. 11–29). New York: Guilford Press.

Williams, K. T. (2007). *Expressive vocabulary test* (2nd ed.). 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.