A Multi-Site Randomized Controlled Trial of the Teacher Study Group Professional Development Program in Vocabulary in First Grade: A Technical Report

2015

Madhavi Jayanthi

Joseph Dimino

Russell Gersten

Mary Jo Taylor

Kelly Haymond

Becky Newman-Gonchar

Instructional Research Group

This project was funded by the U.S. Department of Education (Grant No. R305A090294)

Preferred citation:

Jayanthi, M., Dimino, J., Gersten R., Taylor, M., Haymond, K., & Newman-Gonchar, R. (2015). *A Multi-Site Randomized Controlled Trial of the Teacher Study Group Professional Development Program in Vocabulary in First Grade: A Technical Report.* Los Alamitos, CA: Instructional Research Group.

To download a copy of this document, visit www.inresg.org.

Author Note

The authors would like to acknowledge the contributions of Mengli Song for her advice relating to statistical analyses and Keith Smolkowski for analyzing the data. In addition, the authors would like to thank Pamela Foremski, Jo Ellen Kerr, Samantha Spallone, and Christopher Tran for assisting with the research and preparation of the report.

Table of Contents

Abstract

The purpose of this efficacy study was to examine the impact of the Teacher Study Group (TSG) professional development program in vocabulary, on teacher knowledge, observed teaching practice, and student vocabulary achievement, when implemented with first grade teachers in Title 1 schools. This multi-site cluster randomized trial was implemented in 61 schools from 16 districts in four states (California, Texas, Ohio, and Illinois). The analytic sample included 182 first grade teachers and 1,680 of their students. Significant impacts were found at the teacher level for teacher knowledge ($g = .38$, $p < .05$), the *Teacher-Directed Vocabulary Instruction scale* ($g = .93$, $p < .001$), and the *Interactive Vocabulary Instruction* scale ($g = .47$, $p < .05$). No significant impacts were found at the student level on the *Woodcock-Johnson Oral Vocabulary* and *Reading Vocabulary* subtests and the *GRADE Word Meaning* subtest.

Introduction

Rich, in-depth vocabulary instruction, as articulated in the seminal work of Beck and colleagues (e.g., Beck, McKeown, & Kucan, 2002), as well as in the work of other leading researchers of vocabulary instruction (e.g., Baumann & Kame'enui, 2004; Graves, 2006; Hiebert & Kamil, 2005; Stahl & Nagy, 2006) has been shown to enhance the vocabulary knowledge of a wide range of students across grade levels (e.g., Beck & McKeown, 2007; Coyne, McCoach, Loftus, Zipoli, & Kapp, 2009; Lesaux, Keiffer, Faller, & Kelley, 2010; Schwanenflugel et al., 2010; Silverman & Hines, 2009). This type of rich in-depth vocabulary instruction goes far beyond providing a mere definition or a synonym for a word. Instead, it is characterized by systematic instruction orchestrated by the teacher. Easily comprehensible student-friendly definitions are provided to the class, and the word's meaning(s) are further clarified using both verbal and visual examples (and non-examples), that not only help pinpoint the meaning of the words in various contexts, but also illuminate the subtle nuances and connotations of the words.

Student understanding is further cemented by high-quality extension activities that go beyond just asking students for the definition of the word; they require students apply the word in various contexts through speaking and writing activities. Although experts such as Biemiller (2004) differ on the role of in-depth vocabulary instruction on a small set of words versus more abbreviated instruction in large sets of vocabulary words, most agree that in-depth rich vocabulary instruction can be a beneficial aspect of literacy instruction.

There is a general consensus and understanding in the field that this type of in-depth vocabulary instruction is a rarity in current teaching practice (e.g., Beck & McKeown, 2007; Scott, Jamieson-Noel, & Asselin, 2003). Observational data from several large-scale studies conducted in first and fifth grade classrooms, including our own observational research, support this assertion (Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; James-Burdumy et al., 2009, 2010; Moss et al., 2008).

Most studies focus on specific instructional activities to provide such instruction. Although most provide some level of professional development or training for either the teachers or the research assistants involved in the study, the focus is more on ensuring quality of researcher-developed interventions, rather than assessing the impact of professional development on teachers' routine vocabulary instruction. Very few studies have focused on examining the impact of professional development programs in vocabulary on both teaching practice and student vocabulary outcomes.

To address this dearth in research, the research team has worked on developing and examining the effectiveness of a PD program in vocabulary over the past decade. The first study in this line of research (Gersten et al., 2010) was conducted from 2004 to 2006. It focused on the development of the PD program based on the Beck-McKeown model of in-depth vocabulary instruction and included a relatively small-scale randomized controlled trial involving first grade teachers in 19 schools. To our knowledge, this is the only rigorous study of a PD program in vocabulary that is not linked to a particular curriculum.[1] To extend this line of research, we recently conducted

[1]The PD program that we developed does not include a curriculum and was intentionally not linked to any one curriculum. The goal of the program is to improve teaching practices in vocabulary so that teachers

a replication study with a large-scale sample (2009-2012). We describe this replication study, the findings, and implications in this report.

The Teacher Study Group (TSG) Professional Development Program

The professional development program developed and tested over the past 10 years uses *Teacher Study Groups* as a vehicle for building the quality of teaching practices to conform with the research on vocabulary instruction, regardless of curriculum used. The Teacher Study Group, as we conceptualized it, is similar to other small group professional development models such as lesson study (e.g., Lewis, Perry, Hurd, & O'Connell, 2006) and shares many characteristics with professional learning communities (PLCs) as studied by Borko (2004) and Vescio, Ross, and Adams (2008). In Table 1, we briefly highlight commonalities and differences between Teacher Study Groups and other small-group professional development models. (For additional information, see Gersten et al., 2010). As the table indicates, many features of the TSG are possible features of Lesson Study and PLC PD models, except for the cumulative review feature that no other small-group PD delivery models share with the TSG approach.

---

can use the information presented in the PD sessions to enhance their lessons regardless of the curriculum they use. This is in contrast to other recent studies where the professional development was provided to facilitate teacher grasp of a curriculum material, such as the study by Apthorp et al. (2012) or Schwanenflugel et al. (2010). This is further discussed on page 4.

Table 1

A Comparison of Teacher Study Groups with Other Small-Group PD Delivery Models

	TSG	Lesson Study/PLC
Commonalties		
1. Opportunities for active learning.	Yes	Yes
2. Collective participation of teachers in an array of activities related to lessons to be taught in near future.	Yes	Yes
3. Builds and support collegial support networks.	Yes	Yes
Potential Differences		
4. Aligns with school curriculum.	Yes	Varies
5. Limited to teachers within a grade, within a school.	Yes	Varies
6. Led by a school-based facilitator with strong content and pedagogical knowledge.	Yes	Varies
7. Lesson planning uses research concepts to guide activities.	Yes	Varies
8. Teachers implement planned lessons in classroom.	Yes	Varies but frequently yes
9. Provides a cumulative review of research concepts.	Yes	No
10. Implemented with a semi-structured format.	Yes	Varies

The TSG professional development program in vocabulary was designed to incorporate the "promising practices" of professional development articulated by experts in the field over the past 30 years (as articulated by Birman, Desimone, Porter, & Garet, 2000; Desimone, 2009; Hill, Beisiegel, & Jacob, 2013). These practices include sustained active participation of teachers (e.g., Borko, 2004; Smylie, Mayrowetz, Murphy, & Louis, 2007), coherence with major school curricular goals (Desimone, Garet, Birman, Porter, & Yoon, 2003; Fullan, 2008; Penuel, Fishman, Yamaguchi, & Gallager, 2007), and supportive collegial networks that allow for collective participation and learning (Buysse, Sparkman, & Wesley, 2003; Penuel et al., 2007; Talbert & McLaughlin, 1994). We also designed the professional development program to

incorporate practices focused on effectively translating research into practice, especially use of concrete examples taken from teachers' actual curriculum, so that teachers develop a thorough understanding of both the research concepts and the practical applications of those concepts in their classrooms (e.g., Gersten, Morvant, & Brengelman, 1995; Moats & Foorman, 2008).

The TSG PD program in vocabulary allows teachers to use their own literacy curriculum—be it a core reading series, guided reading using leveled texts, or some combination of the two—and to work with grade level colleagues in developing lessons that include evidence-based vocabulary instruction. The goal of the PD program is to help teachers begin to think about and ultimately to use research-based instructional concepts in their classrooms by integrating the TSG content into their existing curriculum. Therefore, the purpose of the TSG PD program was not to change a district's core curriculum but to *enhance* implementation of that curriculum (Gersten & Brengelman, 1996; Smylie et al., 2007) by using research-based strategies that may not be included in the teacher's guide.

The vocabulary content for the professional development program has been drawn from the research on rich, focused vocabulary instruction (e.g., Baumann & Kame'enui, 2004; Beck et al., 2002; Graves, 2006; Hiebert & Kamil, 2005; Stahl & Nagy, 2006). These research studies emphasize the need for carefully selecting vocabulary words that are essential for understanding the text or that are likely to be used frequently in texts students encounter in the future. Other instructional practices deemed essential include developing easily comprehensible "student-friendly" definitions of new

words, providing concrete examples as well as non-examples of concepts and terms, and developing activities that force students to think about and use the new words in both speaking and writing. These key instructional practices are an integral part of the TSG intervention.

Researchers have examined the effectiveness of these key instructional practices with elementary grade students (e.g., Apthorp et al., 2012; Coyne et al., 2009; Goodson, Wolf, Bell, Turner, & Finney, 2010; Silverman & Hines, 2009). For instance, Apthorp et al. (2012) conducted a randomized controlled trial to examine the effects of a supplemental vocabulary program (*Elements of Reading*) grounded in research-based strategies for students in Grades K, 1, 3, and 4. The intervention program included systematic instruction in 6 to 8 words and multiple exposures and opportunities for use via both oral and written activities. The authors note that the treatment resulted in statistically significant effects across the four grades on researcher-developed measures of knowledge of the words taught (effect size .95 to 1.24). Similarly, Goodson et al. (2010) examined the impact of a supplemental vocabulary program (*PAVEd for Success*) on the vocabulary achievement of nearly 1300 kindergarten students with a randomized controlled trial. Words in this intervention were also taught explicitly and were reinforced through repeated exposures using extension activities and classroom conversations. The authors found a statistically significant impact, effect size of .13 on a standardized test. Data from this new wave of rigorous studies not only provide support for the use of key instructional strategies articulated in the literature but also show

statistically significant outcomes in vocabulary, thus providing an empirical foundation for the vocabulary content delineated in the TSG PD intervention.

Description of the Small-Scale Randomized Controlled Trial of the TSG PD Intervention in Vocabulary (2004-2006)

The first study in this line of research—a small scale multi-site randomized controlled trial—examined the impact of the TSG PD intervention in reading on first grade teachers and their students (Gersten et al., 2010). The study was conducted in 19 schools from three states. Eighty-one teachers and their 575 students constituted the sample. The PD intervention, provided by the research staff at each school site, focused on both comprehension and vocabulary. Since vocabulary is the emphasis of the current study, only the vocabulary portion of PD program and the resulting teacher and student impacts are presented in this paper.

Vocabulary impacts at the teacher level were examined for both teacher knowledge (measured using *Content Knowledge for Teaching Reading Assessment*; Phelps & Schilling, 2004) and observed teaching practice (measured using *Observation Measure for Vocabulary Instruction*). Data were analyzed using a two-level HLM model. Positive and significant impacts favoring the TSG PD condition were found for both teacher knowledge ($g = .73$, $p < .05$) and observed teaching practice ($g = .58$, $p < .01$) using a classroom observational system described in the Method section. Though this study was powered to detect significant impacts on teachers but not students, we report on the marginally significant impact found on the *Woodcock-Johnson Oral Vocabulary* ($g = .44$, $p < .10$), suggesting promise for student vocabulary effects in a larger scale

replication. The effect size was .21 for *Woodcock-Johnson Reading Vocabulary*, which was not statistically significant, but indicative of a potentially promising effect given a larger, more powerful RCT. (See Gersten et al., 2010, for additional details on this study.)

Purpose and Importance of the Replication Study

Given the promise of the TSG PD intervention in the area of vocabulary, we conducted the second study (the focus of this report) to replicate the findings with a much larger sample. Specifically, the purpose of the second study was to assess the impact of the TSG professional development program in vocabulary on (a) teacher knowledge, (b) observed teaching practice, and (c) student vocabulary achievement when implemented with *first grade teachers* in Title I schools.

This replication study serves two important purposes. One, it addresses the much emphasized need for replication studies in the field. Rigorous replication studies help build the knowledge base necessary for making policy-relevant decisions based on sound empirical data, rather than on the premature dissemination of findings from just one study. Replication studies help the field distinguish interventions with reasonably consistent impacts from those with erratic impacts, and ultimately, determine the conditions necessary to predict future impacts of intervention. In essence, this replication study will add to the existing data on the effectiveness of a small-group professional development program in vocabulary. Knowing whether the program works or does not work, or knowing under which conditions it works, will help as decisions are made regarding the type of professional development program to be used.

Method

Setting and Participants

The study took place in 16 districts across four states (California, Ohio, Illinois, Texas). Sixty-two schools were randomly assigned to TSG (n = 31) or control (n = 31) conditions. All schools were Title I elementary schools, serving a diverse population of students. One school chose to leave the study after random assignment, leaving 61 schools in the final analytic sample, and resulting in an attrition rate of 1.64% for the overall sample and 3.23% for the TSG sample. See Table 2.

Table 2

Baseline Demographic Characteristics of Schools in the Analytic Sample

	Total Analytic Sample[a] (N = 61)				
	TSG (n = 30)	Control (n = 31)			
	Mean (*SD*)	Mean (*SD*)	*t*	*df*	*p*
Students Reading at Proficient Level or Higher[b]	62.34 (25.51)	57.32 (25.06)	-0.77	58	.45
English Learners	22.59 (29.65)	26.08 (29.87)	0.46	59	.65
Economically Disadvantaged Students	64.70 (24.32)	64.98 (23.39)	0.05	59	.96
White	49.27 (37.18)	46.59 (38.48)	-0.28	59	.78
Hispanic	35.26 (40.72)	40.57 (41.92)	0.50	59	.62
Black	15.47 (19.31)	12.85 (15.09)	-0.59	59	.56

[a]Total number schools at the time of randomization = 62 (TSG = 31, Control = 31). One TSG school attrited soon after randomization. [b]Reading Proficiency data were available only for 60 schools.

Teacher sample. Two hundred and twenty-six teachers (TSG = 115, Control = 111) comprised the teacher sample at the time of randomization of schools. Of these, one-hundred ninety-one teachers (84.5%) were randomly selected for data analysis to reduce costs.[2] Posttest measures (observations and post implementation surveys) were collected only for these randomly selected teachers. Of these 191 teachers, nine attrited from the study, resulting in a final teacher sample of 182 teachers (Overall attrition = 4.7%; differential attrition = 0.7%). See Appendix A for a pictorial representation of the formation of the teacher analysis sample.

Teacher demographic data are summarized in Table 3. For the analytic sample, 94.68% of the TSG teachers were female, 26.51% possessed a master's degree in education, and 38.55% had coursework beyond a master's degree. Of the control group teachers, 94.32% were female, 28.26% earned a master's degree, and 36.96% had education beyond a master's degree. TSG teachers had, on average, 14.81 years of classroom teaching experience (*SD* = 8.30) and 8.85 years of experience teaching first grade (*SD* = 7.55), whereas control group teachers had 14.87 years of classroom teaching experience (*SD* = 9.09) and 7.93 years of experience teaching first grade (*SD* = 7.03). There were no statistically significant differences between the treatment and control group teachers on any of the variables.

[2]From schools with four or less first grade teachers in the study, all teachers were selected for data analysis (86 teachers were selected in this manner). From schools with five or more first grade teachers in the study, three teachers were randomly selected for data analysis (105 teachers were selected in this manner from a pool of 140 teachers).

Table 3

Baseline Characteristics of the Teacher Analytic Sample

	Total Analytic Sample (N =182)			
	TSG (n = 94)	Control (n = 88)		
	%	%	χ^2 *(df)*	*p*
Gender			0.01(1)	.92
Female	94.68	94.32		
Race/Ethnicity[a]			0.28 (3)	.96
White	62.64	63.10		
Hispanic	27.47	26.19		
Black	5.49	4.76		
Other	4.40	5.95		
Education Level[b]			0.79 (2)	.96
Bachelors	37.78	34.94		
Masters	26.51	28.26		
Beyond MA	38.55	36.96		
Teaching Experience	Mean (*SD*)	Mean (*SD*)	*t (df)*	*p*
Total years of classroom teaching	14.81 (8.30)	14.87 (9.09)	0.04 (180)	.97
Years teaching in Grade 1[c]	8.85 (7.55)	7.93 (7.03)	-0.85 (179)	.40

[a]8 missing/not reported in the analytic sample. [b]8 missing/not reported in the analytic sample. [c]1 missing/not reported in the analytic sample.

Student sample. Our power estimates indicated that a randomly selected sample of eight first grade students per teacher was sufficient to estimate impacts on student outcomes. However, to address potential attrition, we sampled as many as 10 students per teacher. Students were randomly sampled from a list of students that had active consent from their caregivers. In sum, the original randomly selected student

sample included 1811 students. After attrition, the analytic sample included 1,680 first grade students (overall attrition = 7.2%; differential attrition = 2.0%).

The student demographic data are summarized in Table 4. Chi-square analysis revealed a statistically significant difference in the percentage of female students in the TSG group and the control group for analytic samples.[3] Student ethnicity was marginally significant.

Table 4

Baseline Characteristics of the Student Analytic Sample

	Total Analytic Sample (N = 1680)			
	TSG (n = 863)	Control (n = 817)		
	%	%	x^2 *(df)*	*p*
Gender[a]			**5.11 (1)**	**.02**
Female	47.97	53.49		
Race/Ethnicity[a]			2.97 (3)	.09
White	41.83	37.70		
Hispanic	38.12	39.78		
Black	10.89	12.48		
Other	8.81	8.69		
LEP[b]			1.57 (1)	.21
Yes	26.65	23.99		

[a]14 missing/not reported in the analytic sample. [b]LEP = Limited English Proficiency.

We conducted *t*-tests to compare student means on the pre-test reading measures to assess the equivalence of the two groups. As shown in Table 5, only scores on the *Word Identification Fluency (WIF)* pre-test differed significantly for the

[3]Gender was statistically controlled for in the impact analysis.

treatment and control groups, for the analytic sample.[4] There were no other statistically significant differences between TSG and control students at pre-test.

Table 5

Baseline Pretest Scores for the Student Analytic Sample

	Total Analytic Sample (N = 1680)				
	Intervention (n = 863)	Control (n = 817)			
Pretest	Mean (SD)	Mean (SD)	t	Hedges' g	p
WIF Score	10.75 (12.70)	9.60 (11.33)	-1.96	0.10	.05
LNF Score	49.11 (15.55)	48.77 (16.13)	-0.44	0.02	.66
Woodcock-Johnson *Reading Vocabulary*	451.53 (13.65)	450.75 (13.08)	-1.19	0.06	.23
Oral Vocabulary	457.42 (14.17)	457.14 (14.02)	-0.40	0.02	.69
GRADE *Listening Comprehension*	13.51 (2.95)	13.45 (3.02)	-0.43	0.02	.67
Word Meaning	17.22 (6.35)	16.91 (6.31)	-0.99	0.05	.32

Note. *WIF* = *Word Identification Fluency*, *LNF* = *Letter Naming Fluency*, *GRADE* = *Group Reading Assessment and Diagnostic Evaluation*.

Attrition. The school-level overall attrition was 1.6% and differential attrition was 3.1%. At the teacher level, overall attrition was 4.7% and differential was 0.7%. At the student level, overall attrition was 7.2% and differential was 2.0%. These attrition levels are not considered problematic for a randomized controlled trial (What Works Clearinghouse [WWC], 2014).

[4]We statistically controlled for this variable among others in the impact analysis.

Study Design

Our basic design for addressing the research questions was a multi-site cluster randomized trial, where schools were randomly assigned within sites (Donner & Klar, 2000; Shadish, Cook, & Campbell, 2002). We chose this design for several reasons: randomization eliminates selection bias; school-level rather than teacher-level assignment makes contamination less likely; and within-district assignment leads to perfect equivalence on all district characteristics between the two study groups.

As an incentive for participation in the study and to reduce attrition, teachers and literacy personnel facilitating the groups were remunerated for their participation in the study. In addition, all control schools had access to necessary TSG curricular materials and a training webinar at the end of the study.[5]

Teacher Study Group intervention. The TSG program involved 10 interactive sessions held at the school site twice a month from October to April. Each session lasted approximately 75-minutes. Sessions were scheduled with the building principal–either before or after school–to suit the schedules of first grade teachers and not conflict with other PD activities.

The TSG format consisted of small-group meetings (2 to 7 participants per school). Each TSG meeting was conducted in an informal style to allow for open discussion and collaboration among teachers. A 5-phase recursive process (described below) was instituted during each TSG session to provide a common format for the TSG

[5]At the end of the school year, after the conclusion of the study, professional development training was offered to all control schools. Of the 31 control schools, 17 took part in the training.

sessions across facilitators and sites, while leaving room for flexibility to respond to issues or concerns specific to the site or individual teacher.

The scope and sequence of the sessions was based on *Learning How to Improve Vocabulary Instruction through Teacher Study Groups* (Dimino & Taylor, 2009). This book was written based on experiences with an earlier two-year randomized controlled trial conducted to determine the effects of Teacher Study Groups on pedagogy and student achievement in vocabulary and comprehension (Gersten et al., 2010). The content of the PD sessions was adapted from the vocabulary instruction model developed by Beck et al. (2002).

TSG sessions addressed four key topics: (a) selecting words to teach, (b) developing student-friendly definitions, (c) generation of examples, contrasting examples, and concrete representations of word meanings, and (d) other activities to promote multiple meaningful exposures to new words. In addition, sessions towards the end of the program discussed use of context clues to help determine word meaning and activities that extend word learning beyond the reading lesson.

Five-phase recursive process. The five-phase process was repeated during each session. This recursive process included the following components: (a) Debrief, (b) Discuss the Focus Research Concept, (c) Compare Research with Practice, (d) Plan Collaboratively, and (e) Assignment. Participants began by debriefing the lesson they collaboratively planned in the previous session. First, participants described the lesson they taught, discussed how students responded, and shared any changes or adjustments they made while teaching the lesson. A new research concept was

presented during the *Discuss the Focal Research Concept* portion of the session. Participants reviewed, reflected on, and discussed the research concept before proceeding to the *Compare Research with Practice* portion of the session. In that segment of the session, they compared how the focus research concept aligned with the instructional design of their core reading program. Next, participants incorporated the focus research concept into a lesson they collaboratively planned. Finally, participants were given an assignment to complete before the next session. Typically, the assignment required participants to implement the lesson they developed during the session.

Through the consistent use of this recursive process in each session, the vocabulary content was designed to build cumulatively over the TSG sessions. In selected sessions, the focus research concepts that were taught previously were reviewed and practiced. For example, participants applied the information they learned from Session 1 to complete Session 2. Figure 1 provides a summary of the cumulative review provided throughout the program.

The Dimino and Taylor (2009) book provided the facilitator with a specific "game plan" for leading participants through the five-phase recursive process. Each session included session goals, a focus research concept, an overview of the session, and explicit instructions for completing the five phases of the TSG process.

Figure 1

Cumulative Review of the Vocabulary Focus Research Concepts

	Sessions								
	1	2	3	4	5	6	7	8	9
Categories of Natural Context	x	x	x	x	x	x			x
Selecting Words		x	x	x	x	x			x
Student-Friendly Definitions			x	x	x	x			x
Examples Non-examples Concrete Representations				x	x	x			x
Activities to Promote Word Learning					x	x			x
Using Context to Determine Word Meanings							x		
Reviewing & Extending Word Learning								x	

TSG facilitators. The development team created guidelines for choosing facilitators.[6] The guidelines suggested that principals choose individuals who are recognized by the administrator and their colleagues as having expertise in literacy, are regarded as leaders in their schools, and are able to work with adults and develop teachers' knowledge and skills. The school principal chose the TSG facilitator in most instances. Typically, the facilitators were recognized by their administrators and peers as having expertise in literacy and charged with a variety of professional development efforts in reading such as coaching, facilitating trainings, helping teachers interpret data,

[6]To maintain the integrity of the research study, our project staff was divided into two teams. One team consisting of the developers of the TSG PD program was responsible for overseeing all aspects relating to the implementation of the PD program, including developing fidelity checklists, training facilitators, etc. The second team (i.e., the research team) oversaw the research end of the study—that is, data collection and data analysis.

etc. First grade teachers were not allowed to facilitate a group of their fellow first grade teachers. In most cases, the facilitator was school-based, although in some instances one reading specialist served more than one school. A total of 30 (29 female, 1 male) facilitated the 31 groups after they had received training from the research staff. Twenty-one of the 30 facilitators possessed a master's degree. All had classroom teaching experience (*M* = 13 years, *Mdn* = 11.5, range = 2-30); most (*n* = 28) had worked as reading coaches and indicated that they had provided PD in the past.

Facilitator training. Facilitators attended a two-day training conducted by the development team. The training began with an orientation to the research-based vocabulary concepts covered by the TSG. This was followed by a discussion of the purpose and structure of the TSG. The trainers (i.e., the development team) modeled the five-phase recursive process and the facilitators were given an opportunity to practice selected activities as they proceeded through the training. The participants also learned strategies for grouping TSG participants into working pairs or triads, and for monitoring and motivating their teachers as they progress through each session. During the facilitator training, facilitators were issued a digital recording device. They were taught how to record their TSG sessions and upload them onto a secure, password-protected website.

Coaching of facilitators. All TSG facilitators received coaching from teacher researchers, a cadre of retired educators with extensive experience as teachers and/or administrators. Facilitators were instructed to tape record each session and upload the audio recording within 24 hours. Coaches listened to the recording and evaluated the

facilitator's performance using the TSG Session Feedback form. Coaches noted evidence of the facilitator's ability to clearly convey the session goals, adhere to the five-phase process, respond to teachers' comments, questions, or concerns, pace the session, and build rapport with the teachers, as well as the teachers' grasp of the session's content. The form also consisted of seven quality items rated on a 5-point Likert scale and several open-ended questions such as "Were the session goals accomplished? Explain. What were some of the strengths of the session? What were some of the session's weaknesses?"

Coaches used the completed TSG Session Feedback forms to write recommendations for improving future sessions. They also reviewed the content of the next session in order to provide specific guidance on how to apply the recommendations for improvement to the next session. These recommendations were used to provide individualized feedback to facilitators during one-on-one post-session conferences with the facilitators. Coaches began the post-session conference by focusing on the strengths of the sessions followed by specific recommendations for improving future sessions.

All coaches were supervised by the members of the development team. The development team members reviewed the forms completed by the coaches and, if needed, feedback was provided to the coaches prior to the post-session conferences. While the Session Feedback forms provided a basis for the feedback given to the facilitators, the completed forms were not given to the facilitators.

Coach training. All coaches participated in a one-day training offered by the development team. The training began with an explanation of the TSG coaching model and the TSG Session Feedback form. The trainers modeled how to rate a facilitator's performance, describing how they would rate the facilitator and why. Ample practice opportunities and feedback were provided to the coaches during the training session. Guidance was also provided on holding the post-session conference with the facilitators. Coaches were instructed to begin the conference by describing three strengths of the facilitators. They were instructed to rank order the session weaknesses from major to minor and discuss only the top three weaknesses when they addressed recommendations for improving future sessions.

Control condition. The control condition (business-as-usual) constitutes district or state instituted professional development without the TSG component. Teachers in the control condition did not engage in the TSG or have access to the materials made available to teachers in the TSG condition during the course of the study.[7]

Survey of Professional Development Activities

All teachers in the study recorded their professional development activities in reading by completing on-line logs every month. Drawing upon two existing surveys: the *Professional Development Survey* from our previous study (Gersten et al., 2010) and the *Grade 1 Teacher Survey* from the IES Reading First Implementation Study (Moss, Jacob, Boulay, Horst, & Poulos, 2006) we developed *The Survey of Professional*

[7]At the end of the school year, after the conclusion of the study, professional development training was offered to all control schools. Of the 31 control schools, 17 took part in the training.

Development Activities, to acquire a descriptive picture of the professional development activities in the TSG and control conditions. Teachers were asked to provide information on the type of professional development activities made available to them (e.g., coaching, seminars), the amount of time they spent participating in those activities, and instructional content focus of those activities (e.g., vocabulary, comprehension). To substantiate teacher self-report data, we also asked literacy personnel from each school about the school/district-mandated PD activities of their first grade teachers.

Detailed information about the vocabulary professional development that treatment and control teachers attended during the year is summarized in Table 6. All 94 treatment teachers and 49 control teachers (56%) attended PD in vocabulary during the course of the study. The vocabulary PD that the control teachers attended, addressed many of the same topics and activities as the vocabulary PD that treatment teachers attended. However, treatment teachers attended significantly more hours of PD in vocabulary ($M = 12.72$) than control teachers ($M = 2.95$).

Table 6

A Description of Treatment and Control Teachers' PD Activities in Vocabulary over the Course of the Study (Prior Summer and Eight Months of the School Year)

	TSG Teachers (n = 94)	Control Teachers (n = 88)			
			x^2 (*df*)	*t* (*df*)	*p*
Number of teachers who had PD in vocabulary	94 (100%)	49 (56%)	53.02 (1)		< .001
Average time spent doing PD in vocabulary (in hours)	12.72 (3.24)[a]	2.95 (5.77)[a]		14.20 (180)	< .001
Average number of PD activities in vocabulary	5.79 (5.27)[a]	1.22 (1.55)[a]		8.04 (110)[c]	< .001
Number of teachers who reported having PD that covered the following topics:					
Selecting words to teach[b]	94	34			
Student-friendly definitions[b]	94	27			
Providing examples & non-examples[b]	94	24			
Concrete representations[b]	94	23			
Activities to promote word learning[b]	94	38			
Use of context for effective word learning[b]	94	21			
Use of morphology for effective word learning	5	8			
Incidental word learning through listening or reading	17	25			
Student meta-cognitive aspects of learning	6	10			
Teaching use of dictionary, Thesaurus	6	2			
Number of teachers who reported having the following activities during their PD:					
I was required to practice vocabulary strategies I learned and received feedback about my practice[b]	94	9			
I collaborated with colleagues to plan a vocabulary lesson[b]	94	25			
I developed student activities for vocabulary to use in my classroom[b]	94	30			

	TSG Teachers (n = 94)	Control Teachers (n = 88)
I observed demonstrations of vocabulary teaching strategies at a conference	26	28
I observed teachers using the vocabulary strategies taught in the conference session	11	14
I practiced using assessment data to plan vocabulary instruction in the session	4	12
I was required to practice vocabulary strategies I learned but did not receive feedback about my practice	6	11
I led group discussions about vocabulary	3	3
I demonstrated a vocabulary lesson	3	4
Number of teachers who participated in PD with the following format:		
Small group within schools[b]	94	29
Small group	25	38
Short training	28	22
Longer institute	6	9
Coaching	5	12

[a]These numbers represent standard deviation. [b]Relevant to the TSG PD intervention. [c]Levene's test for equality of variances was found to be violated, $F(1,180) = 831.89$, $p < .001$. Due to this violated assumption, a t statistic not assuming homogeneity of variance was computed.

Over half of the treatment teachers (52%) also attended additional PD in vocabulary over and beyond the TSG PD intervention. See Table 7 for a description of the additional PD that treatment teachers attended. Note that for18 treatment teachers, the TSG program replaced professional development that they were required to attend by their school or district. The rest (n = 76) attended the TSG intervention in addition to the professional development required by their school or district.

Table 7

Description of the Additional PD in Vocabulary that Treatment Teachers Participated in Beyond the TSG PD Program

	TSG Teachers (n = 52)
Average time spent doing additional PD in vocabulary (in hours)	1.72 (3.24)
Average number of PD activities	0.81 (1.00)
Number of teachers who reported having additional PD in the following topics:	
Selecting words to teach[a]	24
Student-friendly definitions[a]	22
Providing examples & non-examples[a]	24
Concrete representations[a]	18
Activities to promote word learning[a]	42
Use of context for effective word learning[a]	22
Use of morphology for effective word learning	5
Incidental word learning through listening or reading	17
Student meta-cognitive aspects of learning	6
Teaching use of dictionary, Thesaurus	6
Number of teachers who reported having engaged in the following activities during the PD sessions:	
I was required to practice vocabulary strategies I learned and received feedback about my practice[a]	4
I collaborated with colleagues to plan a vocabulary lesson[a]	11
I developed student activities for vocabulary to use in my classroom[a]	18
I observed demonstrations of vocabulary teaching strategies during the conference session	26
I observed teachers using the vocabulary strategies taught in the conference session	11
I practiced using assessment data to plan vocabulary instruction in the session	4
I was required to practice vocabulary strategies I learned but did not receive feedback about my practice	6
I led group discussions about vocabulary	3

	TSG Teachers (n = 52)
I demonstrated a vocabulary lesson	3
Number of teachers who participated in PD with the following format:	
Small group within schools[a]	14
Small group	25
Short training	28
Longer institute	6
Coaching	5

[a]Relevant to the TSG PD Intervention.

In addition to vocabulary PD, teachers in the study reported receiving PD in other areas of reading (comprehension, decoding, fluency, and phonemic awareness). See Table 8. Both treatment and control teachers received more PD in comprehension than in other areas of reading. Control teachers had more PD than treatment teachers in all four areas of reading, but only the time spent in PD in decoding was significantly more.

Table 8

Amount of PD Received by Treatment and Control Teachers in Areas of Reading Other than Vocabulary over Summer and Eight Months of the School Year

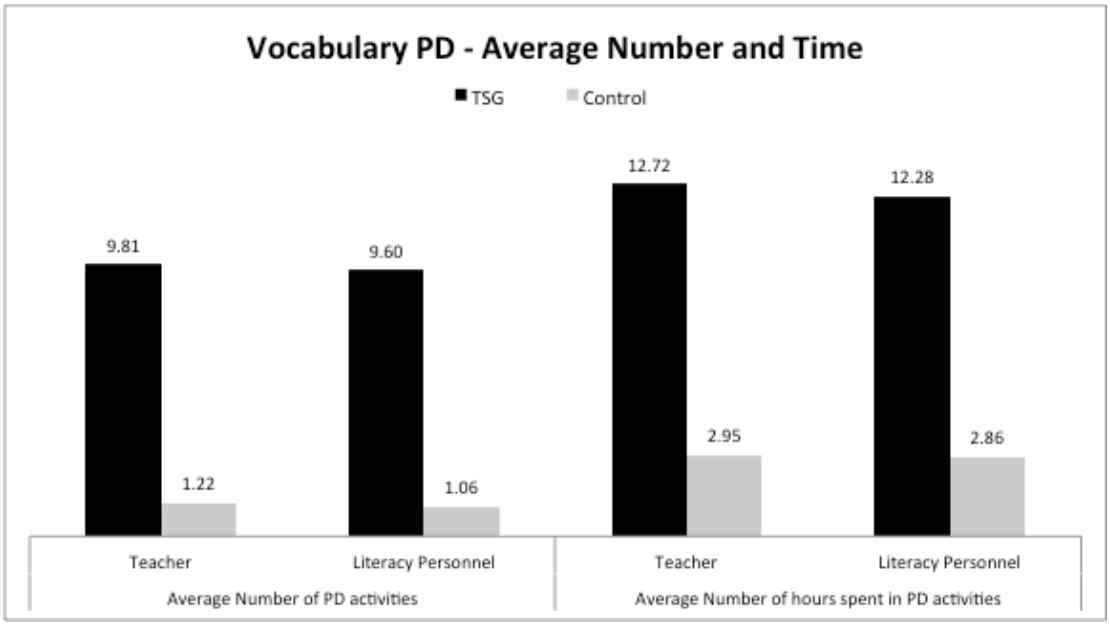
	TSG Teachers (n = 94)	Control Teachers (n = 88)	$x^2(df)$	t (df)	p
Number of teachers who had PD in reading in areas other than vocabulary	74 (78%)	70 (80%)	. 019 (1)		.89
	Hours Mean (*SD*)	Hours Mean (*SD*)			
Comprehension	2.19 (4.61)	3.58 (7.26)		1.55 (180)	.12
Decoding	1.12 (2.14)	1.98 (3.50)		1.99 (142)[a]	.05*
Fluency	1.28 (2.68)	2.11 (3.50)		1.82 (180)	.07~
Phonemic Awareness	1.28 (2.59)	2.09 (3.53)		1.77 (159)[b]	.08~

[a]Levene's test for equality of variances was found to be violated, $F(1,180) = 6.82$, $p < .05$. Due to this violated assumption, a t statistic not assuming homogeneity of variance was computed.
[b]Levene's test for equality of variances was found to be violated, $F(1,180) = 4.37$, $p < .05$. Due to this violated assumption, a t statistic not assuming homogeneity of variance was computed.

Data from the literacy personnel from each school regarding the PD activities of their first grade teachers provides support for the self-report data gathered from teachers. See Figure 2. An independent samples *t*-test revealed that there were no significant differences in the number of PD activities and the number of hours spent in PD activities reported by teachers and literacy personnel.

Figure 2

PD activities reported by teachers and literacy personnel



Assessing Implementation Fidelity

Overall, in the study the TSG PD program was implemented at 31 sites (schools). To assess fidelity of implementation in these sites, all TSG sessions were audio recorded. The research team randomly selected two TSG sessions for the purpose of assessing fidelity. Sessions 4 and 6 were randomly selected from the first half and second half of the TSG program, respectively. The facilitators were not told which audio recordings would be checked for fidelity.

Procedural fidelity. To determine procedural fidelity, checklists that reflected the critical content of Sessions 4 and 6 were developed by the research team. Using these checklists, the research team assessed how well the facilitators adhered to the key procedures for these sessions. Each procedure on the checklist was marked as

observed or not observed. Fidelity was calculated as percentage of procedures implemented (number of procedures observed / total number of procedures [observed and not observed] times 100). See Table 9 for procedural fidelity per session. Mean procedural fidelity for Session 4 was 80.48% (84.62% median; range = 53.85 – 95.24%); For Session 6 it was 94.96 (median = 100%; range = 70 – 100%).

Seven implementations of Session 4 and one implementation of Session 6 had procedural fidelity less than or equal to 75%. Only one implementation of Session 4 had procedural fidelity less than or equal to 60%.

Four Session 4 audio recordings and four Session 6 audio recordings were coded by two researchers in order to assess inter-rater reliability of the fidelity ratings. The mean inter-rater reliability for Session 4 and 6 was 88.39% and 85%, respectively.

Table 9

Fidelity of Implementation

Procedural Fidelity				
	Session 4 (%)		Session 6 (%)	
	Mean (*SD*)	Median (Range)	Mean (*SD*)	Median (Range)
Percentage of key procedures completed	80.48 (10.49)	84.62 (53.85-95.24)	94.96 (9.03)	100.00 (70.00-100.00)
Reliability (% agreement)	88.39 (12.08)	91.07 (71.43-100.00)	85.00 (19.15)	90.00 (60.00-100.00)

Quality of Implementation				
	Session 4 (%)		Session 6 (%)	
Quality Attribute	Mean (*SD*)	Median (Range)	Mean (*SD*)	Median (Range)
Facilitator responds to teachers' comments, questions, concerns	4.34 (0.78)	4.50 (2.00-5.00)	4.60 (0.59)	5.00 (3.00-5.00)
Facilitator paces the lesson so that all parts of the session were covered in sufficient depth	3.94 (0.93)	4.00 (1.00-5.00)	4.47 (0.63)	4.75 (3.00-5.00)
Facilitator uses clarity in conveying session goals	4.40 (0.74)	4.50 (2.00-5.00)	4.75 (0.43)	5.00 (3.50-5.00)
Facilitator adheres to the procedures provided in the manual	4.00 (1.12)	4.50 (1.00-5.00)	4.33 (0.61)	4.00 (3.00-5.00)
Facilitator maintains a positive rapport with teachers	4.84 (0.35)	5.00 (4.00-5.00)	4.82 (0.36)	5.00 (4.00-5.00)
Facilitator's perception of teachers' grasp of the content	3.90 (0.82)	4.00 (2.00-5.00)	4.28 (0.63)	4.50 (3.00-5.00)
Overall rating of facilitator's implementation	4.08 (0.83)	4.00 (2.00-5.00)	4.46 (0.53)	4.50 (3.00-5.00)
Reliability (% agreement, within 1 point of each other)	100.00 (0.00)	100.00 (100.00)	97.14 (6.39)	100.00 (85.71-100.00)

Quality of TSG implementation. The quality of implementations was assessed by rating the facilitators on seven "quality" attributes such as pacing, clarity of the session, and the facilitators' perception of the teachers' grasp of the content and ability to respond to questions, comments, or concerns. All quality items were rated using a 5-point Likert scale, with 1 = low quality, 3 = medium quality, and 5 = high quality. The mean ratings for each quality attribute are displayed in Table 10. Quality ratings in general were lower for Session 4 than for Session 6. The lowest rating for Session 6 was 3, while the lowest rating for Session 4 was 1. For Session 4, facilitators received ratings as low as 1 and 2 for 6 of the 7 quality attributes. The behaviors rated lowest (a rating of 1 or 2) most often were adhering to the procedure provided in the manual (n = 3) and perceiving the teachers' grasp of the content (n = 3), followed by their overall rating for implementing the session (n = 2).

Five Session 4 audio recordings and five Session 6 audio recordings were assessed by two coaches in order for the research team to assess the inter-rater reliability of the quality of implementation ratings. Given the subjectivity of the items on this measure, we defined inter-rater agreement as any two ratings that fall within 1 point of each other on the 5-point Likert scale. The mean inter-rater reliability for Sessions 4 was 100% and for Session 6 was 97.14%.

Teacher and Student Measures used in the Impact Analysis

Teacher measures. Teacher measures for this study include measures of teacher knowledge and observed teaching practice in vocabulary, as well as a measure of teachers' perceptions about the TSG intervention.

1. Measures of Teacher Knowledge. We used the *Content Knowledge for Teaching Reading (CKTR)* assessment (Phelps & Schilling, 2004) as a post-test to measure teacher knowledge in vocabulary. Phelps and Shilling assessed the *CKTR* on several different samples and reported coefficient alphas ranging from .67 to .82. For each sample, estimated IRT reliabilities were above .70. Items in the measure focus on the contextual understanding of vocabulary instruction. Teachers are given classroom scenarios or instructional examples, and are asked questions that relate to instructional decisions based on research-supported practices.

2. Measure of Observed Teaching Practice. We used the *Observation Measure for Vocabulary Instruction (OMVI;* Gersten, Dimino, & Jayanthi, 2007) to assess teaching practice in the area of vocabulary. Two *OMVI* scales, *Teacher-Directed Vocabulary Instruction* and *Interactive Vocabulary Instruction*, were used for confirmatory analysis. These two scales are frequency measures, and the frequency data are recorded in 5-minute intervals. We used a third scale, *Classroom Management and Engagement* for exploratory analysis, as the TSG PD program does not address strategies for classroom management and engagement. The items in this scale were adapted from The Teacher Competency Checklist (Foorman & Schatschneider, 2003). Data on this scale, which includes Yes/No and Likert scale items, are recorded at the end of each observation.

The *OMVI* is well aligned with the extant literature on effective reading instruction (e.g., Anderson, Evertson, & Brophy, 1979; Baumann & Kame'enui, 1991; Beck et al., 2002; Gersten, Fuchs, Williams, & Baker, 2001; Graves, 2006; National Research

Council, 1998). The items reflect two major pedagogical aspects of effective instruction: explicitness of instruction and nature of the interactive instruction (i.e., the amount of scaffolding practice and feedback provided; Ball, 1990; Beck, McKeown, Sandora, Kucan, & Worthy, 1996; National Reading Panel, 2000; Rosenshine & Stevens, 1986). We describe the evolution of the empirically-derived *OMVI* scales and present data on internal consistency, inter-rater reliability, and temporal stability in the next section.

3. Appraisal of the TSG PD program. We administered a survey at the end of the study to gather data on teachers' perceptions and opinions regarding the TSG experience. Our research team developed the survey, *The Professional Appraisal of TSG Survey*, which was also used in our previous study (Gersten et al., 2010).

4. *Measure of Teacher Perceptions of Professional Culture.* To examine the impact of the TSG on teacher perceptions of professional culture in their grade level we used three scales from the Consortium of Chicago Schools Research (2007) surveys—*Quality Professional Development, Uncoordinated Professional Development*, and *Teacher-Teacher Trust*. We combined and adapted two of the professional development scales from the Consortium of Chicago Schools surveys to create a new scale, *The Nature of the Professional Development*. Six of the nine items from the *Quality Professional Development* scale were included to measure teachers' perceptions of how professional development has influenced their teaching and provided them with opportunities to work with their colleagues. We also used two of the three items from the *Uncoordinated Professional Development* scale. These items measure the extent to which professional development activities and topics are

coordinated. Reliability for *The Nature of the Professional Development* scale is .76. The *Teacher-Teacher Trust* scale includes six items and measures the degree to which teachers care for and have mutual respect for each other, and the extent to which they are comfortable sharing their concerns with each other. Scale reliability is .93 for Teacher-Teacher Trust scale.

Student measures. Student measures for this study included measures of entry level reading skills and measures of vocabulary outcomes. The measures used for confirmatory analyses were:

1. Woodcock-Johnson III (WJ) (Woodcock, McGrew, & Mather, 2001). These subtests were administered as posttests to measure vocabulary outcomes. Test-retest reliability for both *Oral Vocabulary* and *Reading Vocabulary* is above .90. *WJ* is an individually administered battery of tests that measures dimensions of reading achievement and related abilities.

2. Group Reading Assessment and Diagnostic Evaluation (GRADE) Word Meaning. The *Word Meaning* subtest was used as a post-test to measure vocabulary outcomes. Test-retest reliability coefficient is above .90. *GRADE* is a diagnostic reading assessment tool (Williams, 2001).

Measures of entry-level skills:

1. Letter Naming Fluency (LNF) (Good & Kaminski, 2002; Kaminski & Good, 1996). This was administered as a pretest to assess student entry-level reading skills. *LNF 6th Edition* has test-retest reliability of .88, and a predictive validity of .65 for reading performance a year later.

2. Word Identification Fluency (WIF) (Fuchs, Fuchs, & Compton, 2004). This was administered as a pretest to assess student entry-level reading skills. The *WIF* has an alternate test-form/stability coefficient of .97 (National Center on Intensive Intervention at American Institutes for Research, n.d.).

3. WJ Reading Vocabulary and Oral Vocabulary. These subtests were administered as pre-tests to measure entry-level skills in vocabulary.

4. Group Reading Assessment and Diagnostic Evaluation (GRADE) Word Meaning and Listening Comprehension. These subtests were given as pre-tests to measure entry-level skills in vocabulary and comprehension. Test-retest reliability coefficients are in the .90 range.

Evolution of the *OMVI* Empirically Derived Scales

The *OMVI* observational measure that was used to collect data on teaching practices included 18 items split among the following three sets of items: (a) *Teacher-Directed Vocabulary Instruction*, (b) *Interactive Vocabulary Instruction*, and (c) *Classroom Management and Engagement*. The first two sets of items are frequency measures, the third set includes items related to classroom management and engagement adapted from Foorman and Schatschneider (2003). In earlier research, we simply treated the full set of items as one scale. Given the larger sample size in this study, we decided to use exploratory factor analysis to empirically generate scales as we had done in our earlier research on reading comprehension (James-Burdumy et al., 2010). In generating these scales, we eliminated items with weak psychometric

properties. In the next section, we detail the manner in which empirically-derived scales were formed for the impact analysis and psychometrically weak items were eliminated.

Item Diagnostics. Table 10 presents item to total correlations for the individual items in the *OMVI*. Based on these item diagnostics, Items 5, 6, 10, and 12 were excluded from the measure as they had either negative or low item to total correlations. All were very low base rate items.

Table 10

Internal Consistency and Cronbach's Alpha

	Correlations with Total
Item Set: Teacher-Directed Vocabulary Instruction	
Item 1: Provide explanation, definition, and/or an example.	.58
Item 2: Elaborate using multiple examples.	.58
Item 3: Elaborate using contrasting examples to pinpoint the definition.	.57
Item 4: Use visuals, gestures, facial expressions, pictures, or demonstrations to determine word meaning.	.55
Item 5: Teach how to use context clues to determine word meanings.	-.01
Item 6: Teach how to use word parts to determine word meanings.	.10
Item Set: Interactive Vocabulary Instruction	
Item 7: Ask students to define words, use words in sentence or state synonyms.	.31
Item 8: Give students opportunity to participate in activities requiring them to demonstrate a deeper understanding of the word.	.46
Item 9: Give students opportunity to use context clues to determine word meanings.	.32
Item 10: Give students opportunity to use word parts to determine word meanings.	.13
Item 11: Further pinpoint the definition by extending or elaborating students' responses.	.78
Item Set: Classroom Management and Engagement	
Item 12: Teacher definition, explanation, and/or example was inaccurate and/or confusing.	-.14
Item 13: Call on about half or more of the students individually.	.38
Item 14: Overall classroom routines.	.87
Item 15: Maximize amount of time available for instruction.	.78
Item 16: Manage student behavior effectively.	.84
Item 17: Students are engaged during the first half of reading block.	.72
Item 18: Students are engaged during the remainder of reading block.	.71

Note. Item to total correlations have not been standardized.

Exploratory Factor Analysis

Next, we conducted an exploratory factor analysis (EFA) in M*plus* (version 7.1) using an oblique rotation called GEOMIN to allow correlated factors. The EFA resulted in only three eigenvalues that exceeded 1.0, suggesting a three-factor solution. See Table 11 for the factor structure and the factor loadings of the three-factor EFA model. As can be seen in the table, there are several items with low factor loadings.

We removed three items with low factor-indicator correlations and ran another EFA. The results from this EFA demonstrated that the data fit the new three-factor EFA model; $\chi^2 = 76.65$, $df = 25$, $p < .0001$; CFI = .95; TLI = .89; RMSEA = .11, 90% CI = [.08, .13]. Table 12 shows the factor structure from the EFA with the abbreviated set of *OMVI* items. Note that one item, *give students opportunity to participate in activities*, cross-loaded on both Factor 1 (*Teacher-Directed Vocabulary Instruction*) and Factor 2 (*Interactive Vocabulary Instruction*). The *Teacher-Directed Vocabulary Instruction* factor correlated with the *Interactive Vocabulary Instruction* factor at $r = .73$ and with the *Classroom Management and Engagement* factor at $r = .18$. The *Interactive Vocabulary Instruction* factor correlated with the *Classroom Management and Engagement* factor at $r = .29$

Table 11

Factor Structure from Three-Factor Exploratory Factor Analysis of OMVI Items

	Factors		
	Factor 1	Factor 2	Factor 3
Items (Factor Indicators)	*Teacher-Directed Vocabulary Instruction*	*Interactive Vocabulary Instruction*	*Classroom Management and Engagement*
Provide an explanation, definition, or example	0.54	0.27	0.10
Elaborate using multiple examples	0.87	0.24	0.09
Elaborate with contrasting examples	0.90	0.13	0.05
Use visuals, gestures, facial expressions, etc.	0.63	0.08	0.07
Ask students to define words, use in sentence	-0.08	0.41†	0.31
Give students opportunity to participate in activities	0.55	0.70	0.25
Give students opportunity to use context clues	-0.04	0.28†	0.18
Teacher pinpoints by extending responses	0.18	1.04[a]	0.32
Call on about half or more students	-0.01	0.18	0.39†
Overall classroom routines (management)	0.08	0.30	0.92
Maximize time available for instruction	0.13	0.32	0.81
Teacher Manages behavior effectively	-0.00[b]	0.28	0.90
Students engagement – first half of reading block	0.03	0.25	0.75
Students engagement – second half of reading block	0.11	0.26	0.75

Note. The factor structure represents the correlation between each indicator and each factor.
†Items with less than 25% overlapping variance with their respective factor (correlation < .50).
[a]The item has a correlation greater than 1.0, which is likely the result of a small estimation error, which is not terribly uncommon.
[b]0.004

Table 12

Factor Structure from Three-Factor Exploratory Factor Analysis of OMVI Items With Poorly Loading Items Removed from the Analysis

	Factors		
	Factor 1	Factor 2	Factor 3
Items (Factor Indicators)	*Teacher-Directed Vocabulary Instruction*	*Interactive Vocabulary Instruction*	*Classroom Management and* Engagement
Provide an explanation, definition, or example	0.54	0.31	0.10
Elaborate using multiple examples	0.87	0.27	0.11
Elaborate with contrasting examples	0.90	0.16	0.07
Use visuals, gestures, facial expressions, etc.	0.63	0.10	0.08
Give students opportunity to participate in activities	0.54	0.74	0.26
Teacher pinpoints by extending responses	0.18	0.98	0.32
Overall classroom routines (management)	0.08	0.29	0.93
Maximize time available for instruction	0.12	0.32	0.81
Teacher manages behavior effectively	-0.00[a]	0.26	0.90
Students engagement – first half of reading block	0.03	0.26	0.74
Students engagement – second half of reading block	0.10	0.26	0.75

[a]-0.002

Next, we created scales using factor-loading weights for all items that demonstrated at least 10% overlapping variance with a factor ($r \geq .32$). We used weighted sum scores, with weights based on factor loadings, because they recognize the relative strength for each item (DiStefano, Zhu, & Mîndrilă, 2009). This allows the

items with the highest loading to contribute most to the factor score. While there is a possibility that the loadings are an artifact of either the specific sample or the chosen extraction or rotation methods, loading-weighted scale scores in general tend to be more stable across samples than true factor scores (Grice & Harris, 1998).

The three loading-weighted scale scores that were created and used in our teacher impact analyses are highlighted in Table 13. Note that one item loaded on two factors so it enters into both of the subscales.

Internal Consistency, Interobserver Reliability, and Temporal Stability of the Three *OMVI* Scales

Data relating to internal consistency, inter-observer reliability, and temporal stability of the three *OMV*I scales are presented in Table 13. The internal consistency **(**Cronbach's α) for *Teacher-Directed Vocabulary Instruction* is .69, for *Interactive Vocabulary Instruction* is .76, and for *Classroom Management and Engagement* is .91.

Table 13

Internal Consistency, Interobserver Reliabilities, and Observation Stability for the Final EFA OMVI Scales

EFA *OMVI* Scales	Internal Consistency (Cronbach's α)	Interobserver Reliability (ICC)	Temporal Stability of Observations (ICC)
Teacher-Directed Vocabulary Instruction	.69	.97	.64
Interactive Vocabulary Instruction	.81	.93	.52
Classroom Management and Engagement	.91	.94	.74

Interobserver reliability. To determine inter-observer reliability, a total of 31 teachers (17% of the sample) were observed simultaneously by two observers. First, we initially calculated mean percent agreement for each item using the percentage agreement method, as it has strong face validity and can be easily interpreted (Stemler & Tsai, 2008). To prevent over inflation and present an objective picture grounded in observed classroom teaching events, we limited our calculation to only the active 5-minute intervals. For an interval to be active, at least one observer had to record data. Thus, intervals with no observed data were excluded from the calculations. We then calculated total number of agreements and disagreements between observer pairs for each item across the entire observation. Finally, we calculated the level of agreement between observer pairs using the following formula: agreements ÷ agreements plus disagreements × 100. The mean percent agreements for the items ranged from 61% to 96% (median = 82%). The one item with a low mean percent agreement (61%) had relatively a low base rate compared to other items. The mean percent agreements for the three scales are as follows: 71% for *Teacher-Directed Vocabulary Instruction*, 82% for *Interactive Vocabulary Instruction*, and 95% for *Classroom Management and Engagement*.

We also calculated intra-class correlation coefficients (ICCs), which are not base rate sensitive or affected by chance (McGraw & Wong, 1996; Shrout & Fleiss, 1979). ICCs were calculated from multi-level models with pairs of observers nested within observation occasions. The reliability model includes two sources of variance estimates: teacher-level variance that constitutes the true variance and the residual error variance

that corresponds to differences between two observers watching the same classroom-teaching situation. ICCs provide an estimate of the proportion of total variance that is accounted for by observer variance. A large ICC indicates that there is very little variation between observers watching the same teaching situation. The ICCs for all three *OMVI* scales are high (above .90).[8]

Temporal stability. Sixty-eight teachers (37% of the sample) were observed twice for the purpose of determining the temporal stability of the observed data. To demonstrate that the frequency and pattern of teaching practices were consistent from day to day, we fit a model with the two observations per classroom nested within each of 68 classrooms.

We used the ICCs here to determine the proportion of the total variance that is accounted for by within-teacher variance across the two observed lessons (e.g., Shoukri, Asyali, & Donner, 2004). High ICCs indicate stable behavior, while lower ICCs imply the need for more observations to obtain a reasonable estimate of teacher behavior. Because ICCs can be interpreted as the average correlation between given pairs of observations for the same teacher across a school year (Shrout & Fleiss, 1979), the stability ICCs reported here are analogous to test-retest reliability estimates.

The ICCs for the three scales are summarized as follows: .64 for the *Teacher-Directed Vocabulary Instruction Scale*, .52 for the *Interactive Vocabulary Instruction Scale*, and .74 for the *Classroom Management and Engagement Scale*. Many of the observed behaviors produced stability ICC greater than .50 indicating moderate to

[8]ICCs can be interpreted along the same guidelines used for kappa (e.g., Landis & Koch, 1977). An ICC of .00–.20 is considered slight reliability; .21–.40 is fair; .41–.60 is moderate; .61–.80 is substantial; and .81–1.00 is nearly perfect.

substantial temporal stability.

Data Collection

All teacher and student data were collected by our cadre of trained observers and data collectors (retired school teachers with a background in reading), with experience in conducting classroom observations in other IES projects (the national evaluations of the earlier TSG study and Reading Comprehension). A demographic survey was administered to teachers and TSG facilitators at the beginning of the study. At the end of the study, all teachers (from both TSG and control condition) were observed once during the entire language/arts block using the *Observation Measure of Vocabulary Instruction* (*OMVI*); 37% of the teachers were observed twice. All other teacher measures (*CKTR, Professional Appraisal of TSG, Nature of the Professional Development* scale, and *Teacher-Teacher Trust* scale) were administered during the last TSG sessions. Facilitators also completed a *Professional Appraisal of TSG* survey at the end of they study.

All pre and post student assessments were administered during independent seatwork to maximize teacher instructional time during the school day. *WJ* was administered individually, while *GRADE* was administered in small groups. Student and school demographic data (LEP status, free and reduced lunch, economically disadvantaged, ethnicity, gender) were gathered from the school databases.

Training for teacher observers. Observers participated in a one and a half-day training session provided by the development team. Each participant received a codebook that included an explanation of both the measure and the rules for coding

instructional behaviors. Training began with a general description of the *OMVI*. In-depth instruction on the measure started with a discussion of the major constructs of effective vocabulary instruction (i.e., explicit instruction, student practice) upon which the observation measure was based.

The trainers operationally defined each item and clarified the rules for coding. To lessen observers' anxiety, coding practice was scaffolded to ease observers into the process. Observers initially viewed and coded short segments of classroom instruction (2-3 minutes) and then proceeded to code longer segments of classroom instruction. During initial practice, each practice instructional segment was first only viewed and not coded. During the second viewing, the observers coded the instruction. During the third viewing, the trainer debriefed participants by stopping the tape each time the teacher earned a tally. The trainer discussed the rationale for the tally, answered questions, and addressed concerns.

The next series of teaching clips were longer in duration. During these coding practices, observers were not given the opportunity to view the film clip before tallying teacher behaviors. However, on the second viewing, participants checked and discussed their coding as discussed above. Towards the end of the training, observers were practicing using 15-minute video segments. After coding each video, the trainer debriefed participants using the procedures described above. As a final step, all participants were assessed to see if they were coding reliably. Participants had to code two 30-minute teaching segments for the purpose of establishing reliability. By the end of each observer training (initial training in Year 1, retraining in Years 2 and 3),

interobserver reliability ranged from 80-91%; that is, 80-91% of the time observers' tallies fell within 1 tally of each other.

Quality control. To provide some initial in-field support, experienced observers were paired with novice observers during the first week of observation. This allowed the observers to code in real time and discuss their codes with another observer to ascertain that they were on target in applying the coding rules.

Training for student data collectors. Student data collectors were trained to administer and score student assessments in a 5-hour training session. The purpose of each assessment and the rules for administering and grading were discussed and modeled. During the training session, participants practiced grading assessment protocols either by viewing videotaped testing sessions or by testing a partner. At the end of the training, accuracy in administration and scoring was checked during mock testing sessions.

Data Analysis Plan: Calculating Treatment Effects on Teacher and Student Outcomes

Given that the data we used to address the research questions are of a nested nature (i.e., students and teachers nested within schools), we used multi-level modeling to perform the main impact analyses. Traditional regression analyses in this case would ignore the dependence among students and teachers nested within the same schools, and as a result lead to underestimated standard errors and potentially wrong conclusions about the TSG's impact. Multi-level models, in contrast, explicitly take into

account the nested data structure, and thus produce properly computed impact estimates and their standard errors (Raudenbush & Bryk, 2002).

For the confirmatory analyses at the teacher level, we examined the impact of the TSG PD program on the two teacher level outcomes—teacher knowledge in vocabulary (measured by *CKTR*) and observed teaching practice in vocabulary (measured by two *OMVI* scales: *Teacher-Directed Vocabulary Instruction* and *Interactive Vocabulary Instruction*). In addition, exploratory analyses were conducted to explore the impact of the TSG intervention on another *OMVI* scale—*Classroom Management and Engagement,* and on two scales of professional culture, *The Nature of the Professional Development* scale and the *Teacher-Teacher Trust* scale. We also examined the interactions between certain moderators (i.e., teacher experience and amount of university level coursework) and the TSG intervention (Jaccard & Turrisi, 2003).

For both confirmatory and exploratory analyses of the TSG's impact on teacher outcomes, we used a two-level model with teachers at Level 1 and schools at Level 2. The model included dummy variables for each randomization block.[9] The intercept of the teacher-level model, which represents average teacher outcome for a given school, is modeled as a random effect at the school level. The randomization blocks were also modeled as random effects. For the exploratory analyses of moderation effects, the

[9]Each randomization block, in all but three instances, equaled a school district. In other words, schools from each district were randomly assigned to treatment or control. However, in two large districts, some schools were sampled during one year, and another set of schools were included during another year of the study. In these instances, the blocks represent the cohorts from those districts. In a third case, a school district had both track and regular schools. In this case, one randomization block contained track schools from the district and another one contained regular schools from the same district.

models were expanded to test moderation by including the moderator and its interaction with condition to test for differential response.

We included years of teaching experience and education level (Master's vs. Bachelor's) as teacher-level covariates. At the school level, we also included four covariates: (a) the percent of students classified as limited English proficient, (b) the percent of minority students, (c) the percent of students who received free and reduced lunch, and (d) the percent of economically disadvantaged students.

For the confirmatory analyses at the student level, we examined the impact of the TSG PD program on student oral vocabulary achievement measured using the *WJ Oral Vocabulary* subtest and on reading vocabulary achievement using two measures, the *WJ Reading Vocabulary subtest and GRADE Word Meaning* subtest. The impact analyses for student outcomes were based on a similar multi-level model, where students are at Level 1 and schools at Level 2.[10] The intercept of the student-level model, which represents average student achievement for a given school is modeled as a random effect at the school level (Level 2). Consistent with the model used for determining teacher level impacts, randomization blocks were modeled as random effects.

[10]The two-level model is more consistent with the sampling design for our study than a three-level model. Our decision is also consistent with the recommendation by Schochet (2008) and Zhu, Jacob, Bloom, and Xu (2012) about how to handle clustering for studies that randomly assign schools to conditions within districts. Even though we believe a parsimonious two-level model is adequate, we also performed the analysis using a three-level model as a sensitivity test and found results were similar. No significant impacts were found on the *WJ Oral Vocabulary* and *Reading Vocabulary* subtests and the *GRADE Word Meaning* subtest.

We included the following student-level covariates: (a) *WJ Oral Vocabulary* pretest, (b) *WJ Reading Vocabulary* pretest, (c) *GRADE Word Meaning* pretest, (c) *GRADE Listening* pretest, (d) *Letter Naming Fluency* pretest, (e) *Word Identification Fluency* pretest, (f) student gender, (g) student Black status, (h) student Hispanic status, and (i) student eligible for free/reduced lunch. At the school level, we included the same covariates used in our teacher model: (a) the percent of students classified as limited English proficient, (b) the percent of minority students, (c) the percent of students who received free and reduced lunch, and (d) the percent of economically disadvantaged students.

We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute, 2009) using restricted maximum likelihood (REML), generally recommended for multilevel models (Hox, 2002). With 61 schools, tests without blocking used 59 degrees of freedom, but this number was reduced to account for interaction terms involving condition and one less than the number of randomization blocks (Murray, 1998).

In all of our models, we standardized the measures to have a mean of zero and standard deviation of one. Therefore, the coefficient for the treatment variable represents the standardized mean difference, that is, the effect size, between TSG and control schools. Following WWC 3.0 guidelines (WWC, 2014) we also computed Hedges' *g* (Hedges, 1981) for each fixed effect.

Results

Teacher outcomes. Results from the multilevel models used to estimate the TSG treatment effects on teacher knowledge of vocabulary instruction and observed teaching practice are presented in Table 14. We used the *Content Knowledge for Teaching Reading (CKTR)* assessment (Phelps & Schilling, 2004) to assess the impact of the intervention on teacher knowledge of vocabulary. The impact was statistically significant, $p = .03$ with an effect size (Hedges'*g*) of 0.38.

Two *OMVI* scales—*Teacher-Directed Vocabulary Instruction* and *Interactive Vocabulary Instruction,* were used to assess impact on teachers' actual day-to-day teaching of vocabulary during reading lessons. The effect size (*g)* for the *Teacher-Directed Vocabulary Instruction* scale was 0.93 and statistically significant ($p <. 001$). The impact was also statistically significant for *Interactive Vocabulary Instruction,* $p = .02$, with an effect size (*g*) of 0.47.

Impact on classroom management and engagement. As an exploratory analysis, we examined the impact of the TSG intervention on teachers' *Classroom Management and Engagement* as measured by the relevant *OMVI* scale. We did not hypothesize any impact in this domain and did not find one (Hedges' $g = 0.03$; $p > .05$).

Impact on teacher perceptions of professional culture. *The Nature of the Professional Development* scale and *Teacher-Teacher Trust* scale were used to measure teacher perceptions of professional culture. Our findings suggest that teachers in the experimental condition perceived the professional developed they received to be significantly higher in quality and cohesion than those in the control group (Hedges' *g* =

0.54; $p < .001$). However, there was no significant difference between groups on the scale measuring teachers' trust and respect for each other.

Student outcomes. We estimated the impact of the TSG intervention on three vocabulary measures—two reading vocabulary measures (*WJ Reading Vocabulary* and *GRADE Word Meaning and one oral vocabulary measure (WJ Oral Vocabulary*). There were no significant impacts on any of the individual measures as noted in Table 15. Effect sizes were all minimal.

Table 14

Teacher Impact Estimated with a Mixed-Model Analysis of Covariance and Random Blocks

		Observed Teaching Practice (*OMVI*)			
	Teacher Knowledge (*CKTR*)	*Teacher-Directed Vocabulary Instruction* (Teacher Explanations)	*Interactive Vocabulary Instruction* (Student Practice)	*Classroom Management and Engagement* (Observer Impressions)	Nature of the Professional Development (Professional Culture)
Fixed Effects					
Intercept	-.18 (.52)	-.17 (.62)	.89 (.63)	1.32* (.55)	.01 (.56)
Condition	.37* (.16)	.85*** (.22)	.46* (.19)	.03 (.19)	.52*** (.13)
Years Teaching	-.01 (.01)	.00 (.01)	-.01 (.01)	-.02~ (.01)	.00 (.01)
Master's Degree	.15 (.17)	-.26* (.12)	-.23 (.15)	-.31~ (.17)	-.05 (.17)
School					
Percentage LEP	.01 (.00)	.01 (.01)	.00 (.01)	.01 (.00)	.00 (.01)
Percentage Minority	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
Percentage FRL	.00 (.01)	.00 (.01)	-.01 (.01)	-.01 (.01)	.00 (.01)
Percentage ECON	.00 (.00)	-.01 (.00)	.00 (.01)	.00 (.00)	.00 (.00)
Random Effects					
Block (Intercept)	.01 (.06)	-.03 (.10)	.02 (.11)	-.02 (.08)	.08 (.08)
Block (Condition)	.01 (.09)	.10 (.22)	-.01 (.18)	.08 (.12)	-.12$^{~}$ (.07)
School (Intercept)	.08 (.11)	.45* (.17)	.40* (.16)	.13 (.11)	.20 (.12)
Residual	.88*** (.12)	.33*** (.04)	.54*** (.07)	.76*** (.10)	.81*** (.11)
ICC	.08	.57	.42	.15	.20
Hedges' *g* (Condition)	0.38	0.93	0.47	0.03	0.54
p-value (Condition)	.03	<.001	.02	.89	<.001

Note. Test of condition was conducted with 55 degrees of freedom. Only the dependent variables (and not the predictors) have been standardized. Table entries show parameter estimates with standard errors in parentheses. Tests of fixed effects (first four rows) used 59 *df* to account for the school as the unit of analysis and the four school-level covariates. ICC = intraclass correlation coefficient *OMVI* = *Observation Measure for Vocabulary Instruction*, *CKTR* = *Content Knowledge for Teaching Reading*, LEP = Limited English Proficient, FRL = Free or Reduced Lunch, ECON = Economically Disadvantaged.

[a]The Benjamini-Hochberg correction resulted in a critical *p*-value of 0.025. The effects remain statistically significant at this critical *p*-value.

†Variance constrained to zero. $^{~}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 15

Impact of the TSG Intervention on Student Oral and Reading Vocabulary

	Oral Vocabulary	Reading Vocabulary	
	WJ Oral Vocabulary	*WJ Reading Vocabulary*	*GRADE Word Meaning*
Fixed Effects			
Intercept	-21.60*** (1.01)	-17.55*** (.90)	-3.18* (1.17)
Condition	.00 (.03)	-.07* (.03)	-.05 (.04)
Student pretest			
WIF	.01* (.00)	.00 (.00)	-.01** (.00)
LNF	.00 (.00)	.01*** (.00)	.01*** (.00)
WJ Reading	.01*** (.00)	.02*** (.00)	-.01* (.00)
WJ Oral	.03*** (.00)	.02*** (.00)	.01*** (.00)
GRADE Word Meaning	.01$^{\sim}$ (.00)	.05*** (.00)	.09*** (.00)
GRADE Listening Comprehension	.05*** (.01)	.03*** (.01)	.03*** (.01)
Gender	-.06$^{\sim}$ (.03)	.06* (.03)	.05 (.04)
Black	-.08 (.06)	-.05 (.05)	-.02 (.07)
Hispanic	.01 (.06)	-.03 (.05)	-.06 (.07)
LEP	-.07 (.05)	.04 (.05)	.21*** (.06)
School			
Percentage LEP	.00 (.00)	.00 (.00)	.00 (.00)
Percentage Minority	.00 (.00)	.00 (.00)	.00* (.00)
Percentage FRL	.00 (.00)	.00 (.00)	-.01$^{\sim}$ (.00)
Percentage ECON	.00 (.00)	.00 (.00)	.00 (.00)
Random Effects			
Block (Intercept)	.01$^{\sim}$ (.01)	.01$^{\sim}$ (.01)	.02* (.01)
Block (Condition)	.00 (.00)	.00 (.00)	†
School (Intercept)	.00 (.01)	.01$^{\sim}$ (.01)	.01 (.01)
Residual	.39*** (.01)	.31*** (.01)	.53*** (.02)
ICC	.01	.03	.01
Hedges' *g* (Condition)	0.00[a]	-0.08	-0.05
p value (Condition)	.98	.03[b]	.23[b]

Note. Test of condition was conducted with 55 degrees of freedom. Only the dependent variables (and not the predictors) have been standardized. Table entries show parameter estimates with standard errors in parentheses. Tests of fixed effects (first four rows) used 59 *df* to account for the school as the unit of analysis and the four school-level covariates. *WJ = Woodcock Johnson* III, *WIF = Word Identification Fluency*, *LNF = Letter Naming Fluency*, *GRADE = Group Reading Assessment and Diagnostic Evaluation*, ICC = intraclass correlation

coefficient, LEP = Limited English Proficient, FRL = Free or Reduced Lunch, ECON = Economically Disadvantaged.
[a].001
[b]The Benjamini-Hochberg correction resulted in a critical p-value of 0.025. Therefore, the effect for *WJ Reading Vocabulary* is not significant.
†Variance constrained to zero. $^{\sim}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Exploratory Analyses

Moderator analysis: Teacher variables. We examined teaching experience and teachers' education level as moderators of the relation between the TSG intervention and teacher knowledge and teacher practice. Analyses revealed that neither were significant moderators.

Relationship between Teacher Outcomes. We conducted correlational analyses to explore the relationship between teacher knowledge and observed teaching practice measured on the *OMVI* as well as the relationships among the three *OMVI* scales. Results are shown below in Table 16. The *Classroom Management and Engagement* scale of the *OMVI* correlated significantly with the other two scales of the *OMVI*, *Teacher-Directed Vocabulary Instruction* ($p < .05$), *Interactive Vocabulary Instruction* ($p < .0001$), and teacher knowledge ($p < .05$). The *Teacher-Directed Vocabulary Instruction* and the *Interactive Vocabulary Instruction* scales of the *OMVI* were significantly correlated ($p < .001$). Teacher knowledge was also marginally correlated with the *Teacher-Directed Vocabulary Instruction* scale ($p < .10$).

Table 16

Correlations Between Teacher Outcomes

Variable	1	2	3
1. Teacher Knowledge (*CKTR*)	--		
2. *Teacher-Directed Vocabulary Instruction*	.13$^{\sim}$	--	
3. *Interactive Vocabulary Instruction*	.02	.73***	--
4. *Classroom Management & Engagement*	.15*	.18*	.29****

Note. Total number of teachers = 182.
$^{\sim}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$.

Relationship between Teacher and Student Outcomes. We also examined the predictive relationship between student posttest achievement scores (i.e., *WJ Reading Vocabulary* and *Oral Vocabulary* subtests and the *GRADE Word Meaning* subtest) and teacher outcomes (i.e., *Teacher Knowledge*, *Observed Teaching Practice scales)*, controlling for student pretest scores. The multilevel model nested students within teachers.

See Table 17. For each student outcome, we fit two models, one with only the pretest values and one with the three predictors. The standardized estimates show the effects when the student level outcomes and the teacher level predictors have been standardized. The Pseudo-R^2 values show the change in teacher-level variance from the unconditional model with no teacher-level predictors and the conditional model with the three teacher-level predictors entered. *Teacher Knowledge* was not a significant predictor for any of the student outcomes. The *Teacher-Directed Vocabulary Instruction* scale significantly predicted student achievement on the *WJ Reading Vocabulary*

subtest ($p < .05$), while the *Interactive Vocabulary Instruction* scale significantly predicted student achievement on both the *WJ Reading Vocabulary* subtest ($p < .01$) and *GRADE Word Meaning* ($p < .05$). The two statistically significant *OMVI* predictors reduced the teacher-level variability estimate from .034 to .031 on the *WJ Reading Vocabulary* subtest, a 9.1% reduction in class-level variance.

District-level analyses. We also explored whether teacher and student level impacts varied by district. No significant impacts were found at the individual districts on either teacher or student outcomes indicating the impact of the intervention did not vary significantly across the districts included in our sample.

Sensitivity Analyses

While our primary impact analyses were based on a random effects model, we also conducted sensitivity analyses which included blocks as fixed effects and one that excluded them from the analysis. As can be seen in Table 18, the results from the analyses that either included blocks as fixed effects or completely excluded blocks varied minimally from our confirmatory random block analyses. This supports Raudenbush and Sadoff (2008), who have shown that when blocking variables are included as random effects, they have no impact on the estimate of the test of condition.

Table 17

Student Literacy Outcomes Predicted by Teacher-Level Variables Controlling for Student Pretest Scores with Students Nested within Teachers

		WJ Reading Vocabulary		*WJ Oral Vocabulary*		*Grade Word Meaning*	
		Unconditional	Conditional	Unconditional	Conditional	Unconditional	Conditional
Fixed Effects	Intercept	-24.50**** (.56)	-24.46**** (.57)	-23.08**** (.56)	-22.99**** (.57)	-1.69**** (.06)	-1.75**** (.10)
	Pretest	.05**** (.00)	.05**** (.00)	.05**** (.00)	.05**** (.00)	.10**** (.00)	.10**** (.00)
	Teacher Knowledge (*CKTR*)		.00 (.11)		-.01 (.11)		.07 (.12)
	Teacher-Directed Vocabulary Instruction		-.01* (.00)		.00 (.00)		-.01 (.00)
	Interactive Vocabulary Instruction		.01** (.00)		.00 (.00)		.01* (.00)
Random Effects	Classroom (Teacher)	.03**** (.01)	.03*** (.01)	.02** (.01)	.02** (.01)	.04*** (.01)	.04*** (.01)
	Residual	.42**** (.02)	.42**** (.02)	.46**** (.02)	.46**** (.02)	.57**** (.02)	.56**** (.02)
ICC		.076	.069	.048	.050	.062	.060
Std Estimates	Teacher Knowledge (*CKTR*)		.000		-.001		.014
	Teacher-Directed Vocabulary Instruction		-.071		-.024		-.054
	Interactive Vocabulary Instruction		.099		.036		.072
Pseudo-R^2	Teacher Level		.091		-.038		.025

$^{\sim}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$

Table 18

Sensitivity Analysis

Outcomes	Block	Coefficient	(*SE*)	*p*	Hedges' *g*
		Teacher Knowledge			
CKTR	Random	0.37	(0.16)	.03	0.38
	Fixed	0.41	(0.17)	.02	0.42
	No Block	0.37	(0.16)	.03	0.38
		Observed Teaching Practice (*OMVI*)			
Teacher-Directed Vocabulary Instruction	Random	0.85	(0.22)	< .001	0.93
	Fixed	0.82	(0.22)	< .001	0.90
	No Block	0.83	(0.20)	< .001	0.91
Interactive Vocabulary Instruction	Random	0.46	(0.19)	.02	0.47
	Fixed	0.43	(0.21)	.02	0.43
	No Block	0.46	(0.20)	.02	0.47
		Student Achievement			
WJ Reading Vocabulary	Random	-0.07	(0.03)	.03	-0.08
	Fixed	-0.07	(0.04)	.06	-0.07
	No Block	-0.07	(0.04)	.11	-0.07
WJ Oral Vocabulary	Random	0.00	(0.03)	.98	0.00
	Fixed	0.00	(0.03)	.94	0.00
	No Block	-0.01	(0.04)	.90	-0.01
GRADE Word Meaning	Random	-0.05	(0.04)	.231	-0.05
	Fixed	-0.06	(0.04)	.143	-0.06
	No Block	-0.04	(0.05)	.420	-0.04

Note. Tests of Condition from Analysis of Fixed Blocks and Without Blocks to Demonstrate the Sensitivity of Condition Effects to the Methods of Incorporating Block in the Analysis.

Teacher and Facilitator Perception on the Usefulness of the TSG PD Program

Teachers' perceptions. Overall 85.71% teachers felt the TSG PD program was helpful in terms of helping them teach vocabulary to first graders and 89.25% said the TSG program was more useful than other professional development they have attended. See Table 19. Most of the teachers indicated that they had learned different ideas in TSG than they had in other vocabulary PD (90.32%). The vast majority of teachers rated each session as useful (82.80 to 96.81% of teachers found them to be useful or very useful).

Of most importance for future refinement of the intervention, the most highly rated sessions were on Selecting Words to Teach and creating Student-Friendly Definitions; 96.81% of teachers found both of them to be useful or very useful. Also of importance is that although 88.04% said the TSG was a good use of their time, only 61.29% of teachers said they would volunteer for another TSG if it were offered again in another area of reading.

Most said the TSG increased their knowledge of teaching vocabulary (98.94%) and improved their skill in teaching vocabulary (91.49%). In fact, 82.80% said they use what they learned in the TSG program frequently and many more planned to use what they learned in the future (97.87%). This may be because they felt what they learned was directly relevant to their teaching (96.81%) and easy to put into practice (85.11%).

Yet most of the teachers commented that it was very difficult to arrange their schedules so that they could attend the TSG sessions. Adding the TSG sessions to their already busy professional and personal schedules was a challenge. These data suggest

that other means of delivery of content such as, for example, a mix of face-to-face sessions on the two critical sessions blended with self-paced online sessions might be a better fit given contemporary context.

Teachers also indicated that they found the *Plan Collaboratively* feature to be most helpful (94.68%). They were decidedly more mixed in their perceptions of the value of debriefing (50.00%) and discussing the research concept (49.84%). Of the teachers who used a basal program, a quarter found that engaging in comparing the research to their teaching was useful (25.81%).

Table 19

TSG Teachers' Satisfaction with the PD Program

Item	Percentage of Teachers Responding
1. Overall, in terms of assisting you to teach vocabulary to first graders, how helpful did you find the TSG professional development program?	Useful or Very Useful[a] 85.71
2. How useful were the following TSG sessions?	Useful or Very Useful[a]
a. Session 1: Words in Context	90.43
b. Session 2: Selecting Words	96.81
c. Session 3: Student Friendly Definitions	96.81
d. Session 4: Examples, Non-examples, & Concrete Representations	89.36
e. Session 5: Activities to Promote Word Learning	93.62
f. Session 6: Cumulative Review I	83.87
g. Session 7: Using Context to Determine Word Meanings	89.36
h. Session 8: Reviewing & Extending Word Learning	84.04
i. Session 9: Cumulative Review II	82.80
3. How often did you implement the skills/ideas presented in the TSG?	Most or All of the Time[b] 82.80
4. If a TSG were offered again at your school in another area of reading (e.g., fluency building, adaptations for ELLs, comprehension, etc.), would you volunteer to be part of it?	Probably or Definitely[c] 61.29
5. How beneficial is the TSG compared with other professional development activities you have attended?	Somewhat Beneficial or More Beneficial[d] 89.25

Item	Percentage of Teachers Responding
6. How helpful were the four features of the TSG program?	Most Helpful or 2nd Most Helpful[e]
a. Debrief: Debriefed experiences in applying the research-based strategies to my teaching.	50.00
b. Discuss the Focus Research Concept: Discussed the research addressed in the readings.	51.06
c. Compare Research with Practice: Reviewed an upcoming lesson and discussed how it does or does not reflect the research principles discussed in the reading.	25.81
d. Plan collaboratively: Designed lessons that incorporate research concepts.	94.68
7. How much do you agree with the following statements?	Agree or Strongly Agree[f]
a. The information presented in the TSG was directly relevant to teaching and learning in my classroom.	96.81
b. The ideas presented in the TSG were easy to put into practice.	85.11
c. The TSG increased my knowledge of how I can teach vocabulary in my classroom.	98.94
d. I was provided with help during the TSG sessions if I was confused.	97.87
e. I learned different ideas in the TSG than in I did in other professional developments that I attended in vocabulary.	90.32
f. My teaching skills in vocabulary have improved as a result of participating in the TSG.	91.49
g. TSG material was presented clearly.	95.74
h. In the future, I plan to use the vocabulary strategies I learned in the TSG.	97.87
i. Attending the TSG was a good use of my time.	88.04
j. I felt comfortable sharing my ideas and concerns during TSG sessions.	95.74

[a]1 = Not at all useful, 2 = Somewhat useful, 3 = Useful, 4 = Very useful. [b]1 = Rarely, 2 = Sometimes, 3 = Most of the time, 4 = All of the time. [c]1 = Definitely not volunteer, 2 = Might volunteer, 3 = Probably volunteer, 4 = Definitely volunteer. [d]1 = Less beneficial, 2 = Somewhat less beneficial, 3 = Somewhat more beneficial, 4 = More beneficial. [e]1 = Most Helpful, 2 = 2nd Most Helpful … 4 = Least Helpful. [f]1 = Strongly disagree, 2 = Disagree, 3 = Agree, 4 = Strongly agree.

Facilitator perceptions. All facilitators thought the TSG professional development helped them teach vocabulary instruction to teachers and that the TSG was more beneficial than other PD. See Table 20. They felt each of the sessions was useful (96.97-100.00%) and, like teachers, they felt the most helpful part of each session was the time the teachers were able to plan collaboratively (90.91%).

All facilitators reported that the TSG program was a good use of their time and that it increased their knowledge and skills for helping teachers improve vocabulary instruction. All facilitators also reported that they planned to use what they learned to help other teachers in the future.

Facilitators thought that the teachers used what they learned and were interested in the new material (both 96.97%). They also reported that the teachers actively participated in the sessions (93.94%). This could be because, as the facilitators reported, the material was directly relevant to the teaching and learning in their school (96.97%).

Most felt that the initial training was adequate in preparing them to facilitate the sessions (84.85%) and that the materials (96.97%) and support (100.00%) provided by the research team helped them facilitate the sessions.

However, approximately two-thirds (63.64%) reported modifying the TSG program in some minor way–either by changing the materials or the timing of the sessions. A couple commented that they used thinking maps more, others created different handouts or posters. Some facilitators extended the duration of the TSG session slightly or rescheduled sessions to fit the needs of the teachers. These changes

did not alter the critical components of the TSG approach, and therefore, they did not impact the facilitator's fidelity of implementation.

Facilitators reported that the most difficult aspects of the TSG program were getting through all the material in the allotted time. Another logistical concern was scheduling sessions given the teachers' other priorities. In fact, most of the comments they made in response to the question about what they would change about the TSG program related to the duration of the sessions and how much they could reasonably do at each session; generally, they wanted more time to cover the material. Several also noted that they found it difficult to work with a group of teachers with different personalities and some had trouble answering teachers' questions accurately given that they had just learned the material themselves.

Table 20

Facilitators' Perceptions on the Usefulness of TSG PD Program

Item	Percentage of Teachers Responding
1. Overall, in terms of assisting you to teach vocabulary instruction to first grade teachers, how helpful did you find the Teacher Study Group (TSG) professional development program?	Helpful or Very Helpful[a] 100.00%
2. How useful were the following sessions of the TSG?	Useful or Very Useful[b]
a. Session 1: Words in Context	100.00%
b. Session 2: Selecting Words	100.00%
c. Session 3: Student Friendly Definitions	100.00%
d. Session 4: Examples, Non-examples, & Concrete Representations	100.00% 100.00%
e. Session 5: Activities to Promote Word Learning	96.97%
f. Session 6: Cumulative Review I	100.00%
g. Session 7: Using Context to Determine Word Meanings	96.97%
h. Session 8: Reviewing & Extending Word Learning	93.94%
i. Session 9: Cumulative Review II	
3. How useful were the materials provided by the research staff for facilitating the TSG?	96.97%
4. How well did the skills/ideas presented in the TSG fit within your school's curricula?	Most or All of the Time[c] 93.94%
5. How beneficial is the TSG compared with other professional development activities?	Somewhat Beneficial or More Beneficial[d] 100.00%
6. If a TSG were offered again at your school in another area of reading (e.g., fluency building, adaptations for ELLs, comprehension, etc.), would you volunteer to facilitate?	Probably or Definitely Volunteer[e] 87.88%
7. How adequate was the initial training provided by the research staff in preparing you to facilitate the TSG?	Adequate or Very Adequate[f] 84.85%
8. How important do you think the ongoing support of the research staff was in helping you facilitate the TSG?	Important or Very Important[g] 93.94%
9. How sufficient was the level of ongoing support provided to you by the research staff?	Sufficient or Very Sufficient[h] 100.00%
10. Please rank the features of the TSG from Most Helpful to Least Helpful.	Most Helpful or 2nd Most Helpful[i]
e. Debrief: Debriefed experiences in applying the research-based strategies to my teaching.	39.39%

Item	Percentage of Teachers Responding
f. Discuss the Focus Research Concept: Discussed the research addressed in the readings.	57.58%
g. Compare Research with Practice: Reviewed an upcoming lesson and discussed how it does or does not reflect the research principles discussed in the reading.	25.00%
h. Plan collaboratively: Designed lessons that incorporate research concepts.	90.91%
11. How much do you agree with the following statements?	Agree or Strongly Agree[j]
a. The information presented in the TSG was directly relevant to the teaching and learning in my school.	96.97%
b. The ideas presented in the TSG were easy for teachers to put into practice.	100.00%
c. The TSG increased my knowledge of how I can assist teachers with vocabulary instruction in their classrooms.	100.00%
d. I think teachers used the vocabulary strategies they learned in the TSG.	96.97%
e. I learned different ideas in the TSG than I did in other professional development activities in vocabulary.	100.00%
f. Teachers seemed interested in the material I presented and the discussions I facilitated during the TSG.	96.97%
g. My ability to help teachers with their teaching skills in vocabulary has improved as a result of facilitating the TSG.	100.00%
h. Teachers actively participated in the TSG.	93.94%
i. Facilitating the TSG was a good use of my time.	100.00%
j. In the future, I plan to provide other teachers with the vocabulary strategies I learned in the TSG.	100.00%
12. Did you modify the TSG professional development program in any way?	Yes
	63.64%

[a] 1 = Not at all helpful, 2 = Somewhat helpful, 3 = Helpful, 4 =Very helpful. [b] 1 = Not at all useful, 2 = Somewhat useful, 3 = Useful, 4 = Very useful. [c] 1 = Rarely, 2 = Sometimes, 3 = Most of the time, 4 = All of the time. [d] 1 = Less beneficial, 2 = Somewhat less beneficial, 3 = Somewhat more beneficial, 4 = More beneficial. [e] 1 = Definitely not volunteer, 2 = Might volunteer, 3 = Probably volunteer, 4 = Definitely volunteer. [f] 1 = Not at all adequate, 2 = Somewhat adequate, 3 = Adequate, 4 = Very adequate. [g] 1 = Not at all important, 2 = Somewhat important, 3 = Important, 4 = Very important. [h] 1 = Not at all sufficient, 2 = Somewhat sufficient, 3 = Sufficient, 4 = Very sufficient. [i] 1 = Most Helpful, 2 = 2nd Most Helpful … 4 = Least Helpful. [j] 1 = Strongly disagree, 2 = Disagree, 3 = Agree, 4 = Strongly agree.

Discussion

This study was intended as a larger scale replication of an earlier randomized controlled trial of a professional development approach. By design, this study used school-level personnel (selected by the principal) to facilitate the sessions, whereas the earlier study used primarily members of the research staff. There were several other important differences between the initial study and the replication. Table 21 outlines these differences.

Because the focus of the initial study was on examining impacts on teacher knowledge and observed teaching practice, the statistical power was quite weak for student outcomes, and they were, in essence, exploratory analyses. (Our a priori power estimate was for a minimal detectable effect size of .35 based on overly optimistic projections of ICC at the school level of .08.) The post hoc power analysis indicated the MDES was actually .57 (since the ICC was actually .22 and the R^2 was lower than anticipated). Nonetheless, the findings in both Oral Vocabulary and Reading Vocabulary were quite promising, with effect sizes of .44 and .21; the Oral Vocabulary impact was marginally significant, with a *p* value less than .10.

In contrast, the current study used more realistic assumptions for power estimates for student outcomes using data from the earlier study as a basis for more refined power estimates. Given a larger budget, we attempted to see whether we could discern impacts on both teacher and student outcomes. We thought, too, that by only focusing on the vocabulary PD (which resulted in larger effects in study one, and seemed to us to be the more developed component), we might replicate the same

promising impacts on vocabulary learning and do so in a fashion that we could detect whether they were statistically significant.

To provide adequate statistical power, the scope of the study was triple that of the earlier study, involving 61 schools and 1,821 students. Although both studies involved only Title I schools, the original study focused almost exclusively on low performing low-income schools with primarily either Hispanic or African American students, whereas the present study included a wider range of ethnic groups and income levels. The scope of the second study was broader involving 16 rather than three districts, and four rather than three states.

Thus, one major question addressed in this replication study was whether the Teacher Study Group model could feasibly be implemented in a relatively large sample of 30 Title I schools in 16 districts in four states across the country. The answer to this question is a resounding "yes." In general, implementation levels for the sessions were reasonably high.

A second major question raised is whether this large-scale implementation would result in impacts in observed teaching performance and teacher knowledge of reading. The answer is "yes". Significant impacts were found at the teacher level for teacher knowledge and teaching practice. Of all the impacts on teaching practice, the impact was strongest on the *Teacher-Directed Vocabulary Instruction* observational scale, which involved aspects of teaching that were more frequently addressed during TSG sessions. In contrast, the *Interactive Teaching* scale demonstrated a somewhat smaller– though still moderately large and statistically significant– effect size. This is not

surprising when considering that interactive teaching is more difficult to alter and the TSG format did not provide the type of role-playing activities or in-class follow-up coaching that might have supported this type of teaching.

A third major research question was whether the TSG would result in significant differences in student vocabulary knowledge, using measures of both oral vocabulary and reading vocabulary. As the reader might recall, the earlier study resulted in promising impacts, especially in Oral Vocabulary, but that effect was what would be considered marginally significant (i.e., $p < .10$). The answer to that question is a resounding no. Impacts were minimal, non-significant and in some cases slightly negative. Again, because different schools and different classrooms utilized different reading programs (and/or level books for guided reading), we used a standardized measures of vocabulary knowledge, (*Woodcock-Johnson Reading Vocabulary and Oral Vocabulary* and *GRADE Word Meaning* test), as we had in the prior study.

Table 21

A Comparison of the Two TSG Studies

	Study 1 2004-2006	Study 2 2009-2012
Focus of PD	• Vocabulary • Comprehension	• Vocabulary
Design	• Multisite cluster randomized trial • School-level random assignment	• Multisite cluster randomized trial • School-level random assignment
Sample	• 3 states: CA, PA, VA • 3 large urban school districts • 19 schools • 81 first grade teachers • 575 students	• 4 states: CA, OH, TX, IL • 16 urban, suburban, rural school districts • 61 schools • 182 first grade teachers • 1,811 students
Attrition	• No attrition at school level • Teacher attrition of 3.57% (loss of 3) • 2% differential • Student attrition of 18.60% (loss of 107) • 3.6% differential	• School attrition of 1.6% (loss of 1) • 3.1% differential • Teacher attrition of 4.7% (loss of 9) • 0.7% differential • Student attrition of 7.2% (loss of 131) • 2.0% differential
Implementation	• TSG groups facilitated mainly by research staff; no formal monitoring or coaching of the facilitators	• TSG groups facilitated by school-identified staff; monitored closely by field staff; ongoing feedback and coaching provided
Teacher Outcomes	• 2–level HLM model (teachers within schools) • Teacher knowledge (*CKTR*)[a] • $g = .73$, $p < .01$ • Observed teaching practice (*OMVI*)[b] • $g = .58$, $p < .01$	• 2–level HLM model (teachers within schools) • Teacher knowledge (*CKTR*)[a] • $g = .38$, $p < .05$ • Observed teaching practice (*OMVI*)[b c] • *Mean* $g = .70$, $p < .001$
Student Outcomes	• 2–level HLM model (students within schools) • Oral Vocabulary (*WDRB*) • $g = .44$, $p < .10$ • Reading Vocabulary (*WDRB*) • $g = .21$, *ns*	• 2–level HLM model (students within schools) • Oral Vocabulary (*WJ*) – Non-significant effect ($g = .00$) • Reading Vocabulary (*WJ*; *GRADE*) – Non-significant effects ($g =$ • -.08 for *WJ* and -.05 for *GRADE*)

[a]*Content Knowledge for Teaching Reading.* [b]*Observation Measure for Vocabulary Instruction.*
[c]Average of *Teacher-directed Vocabulary* and *Interactive Vocabulary Instruction* Scales.

Comparison of Findings from the Two Studies

Table 21 contrasts relevant teacher and student level outcomes across the two studies. The impact on teacher knowledge is smaller in the second study and the impacts on teaching practice differ as well with one larger and the other smaller than the overall impact found in the first study.

Whereas the student outcomes in vocabulary appeared to be moderately large and promising in the earlier study (.44 for reading vocabulary and .21 for oral vocabulary on the *WDRB*), they are now non-significant and small for Oral Vocabulary on the *Woodcock-Johnson* and non-significant for the two measures of reading vocabulary (*Woodcock-Johnson* and *GRADE*).

In this section, we present several possible reasons for the baffling drop in achievement and explore whether the data collected provides any evidence to support these hypotheses. Plausible hypotheses include the following, each of which will be briefly discussed: (a) differing demographics of student participants, (b) quality of implementation and different level of personnel facilitating the PD program across studies, and (c) changing context for teaching and PD for first grade vocabulary and reading instruction over the five year difference between the start of the first and second studies.

Differing Demographics of Student Participants.

Table 22 contrasts the student demographic data from the two studies, presenting median school-level demographics for participating schools in both studies. As can be seen, the two samples are quite different.

Table 22

Differences in School Samples

	Study 1 (2004–2006) (Median)	Study 2 (2009–2012) (Median)
Reading Proficiency	15%	68%
Limited English Proficiency (LEP)	26%	8%
Free/Reduced Lunch	92%	56%
Race/Ethnicity		
Black	55%	5%
Hispanic	45%	13%
White	1%	56%
Minority	99%	44%

It is obvious from Table 22 that the demographics are strikingly different. In study 1, only 15% of the students in each school were deemed proficient in reading, whereas it is 68% percent in study 2. Although study 2 included only Title I schools, in this case 56% of students received free or reduced lunch, as opposed to 92% in Study 1. Similarly, 44% of the students in the schools were from ethnic minority groups as opposed to 99% in study one.

We speculated that perhaps this type of PD led to outcomes in first grade more easily and consistently when students' needs in literacy (and in all likelihood vocabulary) were greater. To explore this hypothesis, we conducted secondary analyses to explore whether there were differences across districts, depending on SES. No such differences were found. However, when the analysis focused on students who scored less than 25 on the *LNF* (i.e., students considered at-risk for reading difficulties), there was a marginally significant impact on the standardized measure of oral vocabulary, $t(100) =$

1.68, $p < .10$; $g = 0.33$. It should be noted, however, that only 6% of the sample scored low enough on the *LNF* pretest to be considered at-risk, which greatly restricted the sample of students. Given these findings, future research should examine the effectiveness of the TSG PD specifically for students with weak vocabulary knowledge and relevant pre-literacy skills.

Quality of Implementation and Nature of Personnel Implementing the PD.

In the first study, many schools only agreed if TSG sessions were on Saturdays or held during teacher prep time (which was only 30 min). Two of the three sites insisted on breaking the 75-minute session into 2-3 brief segments or conducting multiple sessions as daylong marathons; neither of which was ideal for implementation.

In contrast, for the current study, each school committed to providing a 75-minute block of time for these biweekly meetings (on occasion schedules were shifted due to other school events). Schools were only considered as participants if they agreed to the uninterrupted 75-minute block of time for professional development. Implementation was far more consistent across sites. The lesson guides were also far more scripted for this implementation (Dimino & Taylor, 2009).

Although implementation was monitored in both studies, in the current study, sessions were monitored more closely and coaches provided feedback. In the earlier study, however, facilitators had a stronger grasp of the research literature on the topic of vocabulary instruction. The first study was implemented by members of the research staff, all of whom had strong backgrounds in reading research, and some of whom helped in the early conceptualization of the PD approach. In the current study, the

facilitators were school personnel, chosen by the principal to facilitate the group. One might assume the replication study would produce lower quality of implementation, but there was no evidence of this in our analyses of the coaching feedback.

Changing Context

The first study was conducted in 2004-05 and 2005-06 school years during the early years of the *Reading First* program. At that time, the major emphasis of most *Reading First* professional development activities was on teaching students to read– i.e., building phonemic awareness, decoding skills, and fluent word and then sentence reading. The emphasis did evolve during the latter years of *Reading First* to include vocabulary and comprehension, but not at the time of the first study.

Most schools emphasized adherence to a core reading curriculum. Students were moved through the core curriculum at a reasonable pace to ensure access to grade-level material for all students. Also, screening and progress monitoring were used to ascertain students who required additional intervention. Thus teachers were asked to comply with a great many procedures and policy– adherence to a core curriculum, regular use of progress monitoring and universal screening, and adherence to pacing schedules set up by the district.

Reading First provided a good deal of funding to reach these goals and in most states, a good deal of the funds went into supporting literacy coaches. Typically they focused heavily on adherence to the procedural demands of *Reading First*. To counterbalance that effort, we intentionally designed TSG to be professional in nature (as opposed to procedural- and compliance-oriented), and to include a good deal of

collegial interaction around professional issues related to enhancing instruction (Desimone, 2009). Thus, we did not include the visitations and coaching sessions, which were, if anything, overdone with some of the *Reading First* implementation. We also chose to emphasize vocabulary and comprehension because these were not areas that *Reading First* PD was emphasizing at the time.

By the time of the second study, *Reading First* was over, and the schools in our study were allowing a good deal more flexibility in terms of adherence to the teacher's guides that accompanied the core reading series and, in some cases, the use of guided reading as opposed to a core reading series. Most states had included at least some professional institutes that highlighted vocabulary and comprehension instruction.

This context may have been less conducive to TSGs demonstrating an impact on student vocabulary instruction; although no clear mechanism comes to mind. One might hypothesize that teachers no longer needed such intensive work in vocabulary instruction, yet the control group observation scores do not suggest this to be true. Nor does the fact that the TSG program, again, demonstrated significant impacts in both teaching practice and teacher knowledge.

Summary and Conclusions

This study demonstrated that the TSG model used in the two studies led to replicated effects on teaching practice are aligned with contemporary views of the research base on vocabulary instruction (e.g., Baumann & Kame'enui, 2004; Beck & McKeown, 2007). The TSG approach also led to significant impacts on teachers'

knowledge of evidence-based vocabulary instruction. Thus, in many ways, the replication did reach its goals.

There was no discernible impact, however, on student vocabulary knowledge, at least as assessed by two standardized measures of reading vocabulary and one measure of oral vocabulary. Elleman, Lindo, Morphy, and Compton (2009) note that most positive significant impacts in vocabulary growth are found on researcher-developed measures. Mean impact on standardized measures of vocabulary, measures that include many more words than those actually taught–and perhaps no words actually taught, was found to be an effect size of .10. Therefore, the findings of no significant impact on standardized vocabulary measures should not be surprising, especially because the PD intervention had, at best, an indirect effect on students, unlike many of the scripted vocabulary interventions that occur in the research literature. This, of course, does not explain the presence of promising impacts on both Oral and Reading Vocabulary on the standardized *WDRB* measures in the first study.

We still would like to urge districts to consider use of PD such as the Teacher Study Group model because it (a) is sustained work on a crucial instructional topic, (b) does not have the "top down" feeling of many of the trainings and institutes that teachers participate in, and (c) treats teachers as professionals who can contribute ideas and learn from each other. Our original conception of the PD model, as one that is far more professional in tone and far less directive and condescending than many approaches has not altered over the years of the two studies. Note too that the Teacher Study Groups did have a well-sequenced, coherent set of activities to simultaneously

build knowledge and skill in vocabulary instruction and thus, in many ways, is quite different from some of the PD approaches found in other studies (e.g., Garet, Porter, Desimone, Birman, & Yoon, 2001).

References

Anderson, L., Evertson, C., & Brophy, J. (1979). An experimental study of effective teaching in first-grade reading groups. *The Elementary School Journal, 79*(4), 193–223. Retrieved from http://www.jstor.org/stable/1001250

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a supplemental vocabulary program on word knowledge and passage comprehension. *Journal of Research on Educational Effectiveness*, *5*(2), 160–188. doi:10.1080/19345747.2012.660240

Ball, D. L. (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis, 12*(3), 247–259. doi:10.3102/01623737012003247

Baumann, J. F., & Kame'enui, E. J. (1991). Research on vocabulary instruction: Ode to Voltaire. In J. Flood, D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 604–632). Upper Saddle River, NJ: Merrill/Prentice Hall.

Baumann, J. F., & Kame'enui, E. J. (Eds.). (2004). *Vocabulary instruction: Research to practice*. New York, NY: Guilford Press.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal*, *107*(3), 251–271. Retrieved from http://www.jstor.org/stable/10.1086/511706

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford Press.

Beck, I. L., McKeown, M. G., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text.

The Elementary School Journal, 96(4), 385–414. Retrieved from http://www.jstor.org/stable/1001863

Biemiller, A. (2004). Teaching vocabulary in the primary grades: Vocabulary instruction needed. In J. F. Baumann & E. J. Kame'enui (Eds.), *Reading vocabulary: Research to practice* (pp. 28–40). New York, NY: Guilford Press.

Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership*, *57*(8), 28–33.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3–15. doi:10.3102/0013189X033008003

Buysse, V., Sparkman, K. L., & Wesley, P. W. (2003). Communities of practice: Connecting what we know with what we do. *Exceptional Children*, *69*(3), 263–277.

Consortium on Chicago School Research. (2007). Survey of Chicago public schools: Elementary school teacher edition. Retrieved from http://ccsr.uchicago.edu/downloads/66312007_elem_teacher_codebook.pdf

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli Jr, R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth versus depth. *The Elementary School Journal*, *110*(1), 1–18. Retrieved from http://www.jstor.org/stable/10.1086/598840

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. doi:10.3102/0013189X08331140

Desimone, L., Garet, M. S., Birman, B. F., Porter, A., & Yoon, K. S. (2003). Improving teachers' in-service professional development in mathematics and science: The role of postsecondary institutions. *Educational Policy*, *17*(5), 613–649. doi:10.1177/0895904803256791

Dimino, J., & Taylor, M. J. (2009). *Learning how to improve vocabulary instruction through Teacher Study Groups*. Baltimore, MD: Brookes Publishing.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1–11. Retrieved from http://pareonline.net/pdf/v14n20.pdf

Donner, A., & Klar, N. (2000). Cluster randomization trials. *Statistical Methods in Medical Research, 9*(2), 79–80. doi:10.1177/096228020000900201

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (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), 1–44. doi:10.1080/19345740802539200

Foorman, B. R., & Schatschneider, C. (2003). Measurement of teaching practices during reading/language arts instruction and it's relationship to student achievement. In S. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching & learning* (pp. 1–30). Baltimore, MD: Brooks Publishing.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly, 27*(4), 216–227. doi:10.2307/1593674

Fullan, M. (2008). *The six secrets of change*. San Francisco, CA: Jossey-Bass.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Gersten, R., & Brengelman, S. U. (1996). The quest to translate research into classroom practice the emerging knowledge base. *Remedial and Special Education*, *17*(2), 67–74. doi:10.1177/074193259601700202

Gersten, R., Dimino, J., & Jayanthi, M. (2007). Towards the development of a nuanced classroom observational system for studying comprehension and vocabulary instruction. In B. Taylor & J. Ysseldyke (Eds.), *Educational interventions for struggling readers* (pp. 381–425). New York: Teachers College Press.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J., & Santoro L. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal, 47*(3), 694–739. doi:10.3102/0002831209361208

Gersten, R., Fuchs, D., Williams, J., & Baker, D. (2001). Teaching reading comprehension strategies to students with learning disabilities. *Review of Educational Research, 71*(2), 279–320. doi:10.3102/00346543071002279

Gersten, R., Morvant, M., & Brengelman, S. (1995). Close to the classroom is close to the bone: Coaching as a means to translate research into classroom practice. *Exceptional Children, 62*(1), 52–66.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills (6th ed.)*. Eugene, OR: Institute for the Development of Educational Achievement.

Goodson, B., Wolf, A., Bell, S., Turner, H., & Finney, P. B. (2010). *The effectiveness of a program to accelerate vocabulary development in kindergarten* (VOCAB) (NCEE 2010-4014). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_20104014.pdf

Graves, M. F. (2006). *The vocabulary book: Learning & instruction*. New York, NY: Teachers College Press.

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research, 33*(2), 221–247. doi:10.1207/s15327906mbr3302_2

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128.

Hiebert, E. H., & Kamil, M. L. (Eds.). (2005). *Teaching and learning vocabulary: Bringing research to practice.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476–487. doi:10.3102/0013189X13512674

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression series: Quantitative applications in the social sciences*. London: Sage.

James-Burdumy, S., Deke, J., Lugo-Gil, J., Carey, N., Hershey, A., Gersten, R., . . . Haymond, K. (2010). *Effectiveness of selected supplemental reading comprehension interventions: Findings from two student cohorts* (NCEE 2010-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20104015/pdf/20104015.pdf

James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., . . . Faddis, B. (2009). *Effectiveness of selected supplemental reading comprehension interventions: Impacts on a first cohort of fifth-grade students* (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20094032/pdf/20094032.pdf

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*(2), 215–227.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. Retrieved from http://www.jstor.org/stable/2529310

Lesaux, N. K., Keiffer, M. J., Faller, S. E., & Kelley, J. G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, *45*(2), 196–228. doi:10.1598/RRQ.45.2.3

Lewis, C., Perry, R., Hurd, J., & O Connell, M. P. (2006). Lesson study comes of age in North America. *Phi Delta Kappan*, *88*(4), 273–281. Retrieved from http://ww.lessonresearch.net/LS_06Kappan.pdf

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

Moats, L. C., & Foorman, B. R. (2008). Literacy achievement in the primary grades in high poverty schools: Lessons learned from a five-year research program. In S. B. Neuman (Ed.), *Educating the other America* (pp. 91–111). Baltimore, MD: Brookes Publishing. Retrieved from http://www.soprislearning.com/docs/librariesprovider3/other-products-items/letrs_neuman_ch_5.pdf

Moss, M., Fountain, A. R., Boulay, B., Horst, M., Rodger, C., & Brown-Lyons, M. (2008). *Reading First implementation evaluation final report*. Cambridge, MA: Abt Associates, Inc. Retrieved from http://files.eric.ed.gov/fulltext/ED504204.pdf

Moss, M., Jacob, R., Boulay, B., Horst, M., & Poulos, J. (2006). *Reading first implementation evaluation: Interim report*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and

Program Studies Service. Retrieved from http://files.eric.ed.gov/fulltext/ED492928.pdf

Murray, D. M. (1998) *Design and analysis of group randomized trials*. New York, NY: Oxford University Press.

National Center on Intensive Intervention at American Institutes for Research. (n.d.). *Curriculum-based measurement in reading (CBM-R)*. Retrieved from http://www.intensiveintervention.org/chart/progress-monitoring/12825

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific literature on reading and its implications for reading instruction*. Washington, DC: The National Institute for Literacy.

National Research Council. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), 921–958. doi:10.3102/0002831207308221

Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *The Elementary School Journal, 105*(1), 31–48. Retrieved from http://www.jstor.org/stable/10.1086/428764

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

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*(2), 138–154. doi:10.1080/19345740801982104

Rosenshine, B., & Stevens, R. (1986). *Teaching functions. Handbook of research on teaching*. New York, NY: MacMillan Publishing.

SAS Institute. (2009). SAS/STAT (Version 9.2) [Data Analysis Software]. Cary, NC: Author.

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

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. doi:10.1080/1086296X.2010.503551

Scott, J. A., Jamieson-Noel, D., & Asselin, M. (2003). Vocabulary instruction throughout the day in twenty-three Canadian upper-elementary classrooms. *The Elementary School Journal, 103*(3), 269–286. Retrieved from http://www.jstor.org/stable/1002272

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research, 13*(4), 251–271. doi:10.1191/0962280204sm365ra

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. doi:10.1037/0033-2909.86.2.420

Silverman, R., & Hines, S. (2009). The effects of multimedia-enhanced instruction on the vocabulary of English-language learners and non-English-language learners in

pre-kindergarten through second grade. *Journal of Educational Psychology, 101*(2), 305–314. doi:10.1037/a0014217

Smylie, M. A., Mayrowetz, D., Murphy, J., & Louis, K. S. (2007). Trust and the development of distributed leadership. *The Journal of School Leadership, 17*(4), 469–503.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks, CA: Sage.

Talbert, J. E., & McLaughlin, M. W. (1994). Teacher professionalism in local school contexts. *American Journal of Education, 102*(2), 123–153. Retrieved from http://www.jstor.org/stable/1085719

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education 24*(1), 80–91. doi:10.1016/j.tate.2007.01.004

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Williams, K.T. (2001). *Group reading assessment and diagnostic evaluation*. Shoreview, MN: Pearson AGS Globe.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2011). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis, 34*(1), 45–68. doi:10.3102/0162373711423786.

Appendix A

Figure A1 – Formation of the Randomly Selected Teacher Sample