An Investigation of the Impact of the Teacher Study Group as a Means to Enhance the Quality of Reading Comprehension and Vocabulary Instruction for First Graders in Reading First Schools: Technical Report

(2009)

Russell Gersten
Instructional Research Group

Joe Dimino
Instructional Research Group

Madhavi Jayanthi
Instructional Research Group

Jimmy Kim
Harvard Graduate School

Lana Santoro
Instructional Research Group

This Teacher Quality Project was funded by U.S. Department of Education (Award No. R305M030052).

Preferred citation:

Gersten, R., Dimino, J., Jayanthi, M., Kim, J., & Santoro, L. (2009). *An investigation of the impact of the Teacher Study Group as a means to enhance the quality of reading comprehension and vocabulary instruction for first graders in Reading First schools: A Technical Report.* Los Alamitos, CA: Instructional Research Group.

Table of Contents

Abstract

Randomized field trials were used to examine the impact of the Teacher Study Group, a professional development model, on first grade teachers' reading comprehension and vocabulary instruction, their knowledge of these areas, and on the corollary comprehension and vocabulary achievement of their students. The multi-site study was conducted during the 2004-2005 and 2005-2006 school years in three large urban school districts from three states: California, Pennsylvania, and Virginia. A total of 81 first grade teachers and their 468 students from 19 Reading First schools formed the analytic sample in the study. Classrooms observations of teaching practice showed significant improvements in TSG schools: TSG teachers scored .86 standard deviations higher on the comprehension measure and .58 standard deviations higher on the vocabulary measure than control teachers. TSG teachers also significantly outperformed control teachers on the teacher knowledge measure of vocabulary instruction (ES = .73). Confirmatory analysis of student outcomes focused on passage comprehension, reading vocabulary, and oral vocabulary; of these only the latter was marginally significant (.44). Implications for teacher practice and future research are discussed.

Introduction

For almost two decades, research on professional development has suggested that commonly used one or two day trainings, that is, one-shot in-services, are ineffective (Goldenberg & Gallimore, 1991). Large-scale case study research (e.g., Huberman & Miles, 1984) and survey research (Garet, Porter, Desimone, Birman, & Yoon, 2001) suggest that teachers value continuous coherent training opportunities that integrate teacher learning into daily classroom instruction. Fundamentally absent in the short term, and typically large scale, professional development workshops are opportunities for teachers to collectively reflect on the relevance of something before they learn it, obtain new knowledge in real contexts, and connect new ideas and skills to their already diverse teaching experiences (Knowles, Holton, & Swanson, 2005). Although professional development that contains solid conceptual and theoretical material tends to be viewed more positively, teachers crave training opportunities that contain concrete examples relating to their current curricula (McLaughlin, 1990).

In recent years there has been an increased interest in the use of the Teacher Study Groups (TSG) or the Teacher Work Group as an approach to professional development (Carroll, 2005; Lambert, 2002; Meyer et al., 1998; Murphy, 1992; Saunders et al., 2001). The Teacher Study Group intervention is a result of ongoing research on professional development and program implementation conducted by the project director over the past 30 years (Gersten, Morvant, & Brengelman, 1995; Gersten & Brengelman, 1996; Gersten & Woodward, 1990). This approach to professional development is an attempt to orchestrate the major trends in professional development

research – linkage to core curriculum, concreteness, establishment of collegial networks, and ongoing related activities – into a feasible model for use in elementary schools. The goal of the TSG is to help teachers begin to think about and ultimately to use research-based instructional strategies in their classrooms by integrating the TSG content into their existing curriculum. Therefore, the purpose of the TSG is not to change a district's core curriculum, but to *enhance* implementation of that curriculum (Gersten & Woodward) by using research based strategies that may not be included in teachers' guides.

One of the primary limitations with recent research on the TSG is the limited number of high-quality studies (e.g., the use of randomized control trails). More common are discussions of design experiments (Tichenor & Heins, 2000), analytic reviews of research (Lewis, Perry, Hurd, & O'Connol, 2006), discussions of district-implemented professional development with lesson study components (Blum, Yocom, Trent, & McLaughlin, 2005; Lewis et al.), and guidelines for implementing TSGs based on collaborative university-school field research (Fernandez & Chokshi, 2002; Watanabe, 2002).

Given the current emphasis on evidence-based instruction, there is a strong need for systematically evaluating the relative effectiveness of the TSG model. In this study, we examined the impact of the Teacher Study Group on teacher knowledge, observed teaching practice, and student vocabulary and comprehension achievement, when implemented with first grade teachers in Reading First schools. Specifically, our research questions were (a) What was the impact of the TSG on teacher knowledge

and teacher practice? and (b) What was the impact of the TSG on student vocabulary and comprehension outcomes?

Method

Setting and Participants

The multi-site study was conducted in three large urban school districts from three states: California (CA), Pennsylvania (PA), and Virginia (VA). A total of 19 Reading First schools were involved in the study (10 TSG, 9 control). Our initial teacher sample included 84 first grade teachers (40 TSG, 44 control); however, three teachers (1 TSG, 2 control) dropped out of the study for a variety of reasons: family problems, illness, and leaving the school district. Our final analytic teacher sample consisted of 81 teachers. Seven students were randomly selected from each class to examine the impact of the Teacher Study Group. Our initial student sample included 575 students[1] (273 TSG, 302 control), with mobility issues resulting in a final analytic sample of 468 students (217 TSG, 251 control).

The teacher demographic data are summarized in Table 1 (Note: Teacher demographic data by state are provided in Appendix A.) Of the 40 first grade teachers in the TSG group, only three teachers were male. Fourteen had a Master's level degree in an education-related field. On average, TSG teachers had 11.66 years of classroom teaching experience (SD = 9.69; range 1 - 31 years) and specifically, 5.45 years of experience teaching first grade (SD = 6.12; range 0 - 23 years). Similarly, of the 44 first grade teachers from the control group, four teachers were male and 19 teachers had a

[1] Due to oversampling

Master's level degree in an education-related field. On average, the control group teachers had 9.74 years of classroom teaching experience (SD = 9.80; range 0 - 36 years) and 4.35 years of experience teaching first grade (SD = 6.00; range 0 - 32 years).

An independent samples t-test indicated that TSG and control teachers did not differ significantly in their years of classroom teaching experience and in their number of years teaching first grade. A chi-square test on teachers' educational degree revealed a marginally significant difference between TSG and control groups.

Table 1

Teacher Demographic Data

		TSG # of Teachers	Control # of Teachers
Initial sample		40	44
Analytic sample		39	42
State	CA	25	30
	PA	10	10
	VA	5	4
Gender	Male	3	4
	Female	37	40
University Training	Bachelors	40	44
	Masters	14~	19
	Post Masters	6	15
Certification	Elementary	40	44
	Reading Specialist	0	1
	Administrative	1	1
	Other	6	8
Total number of Years of		X (SD)	X (SD)
	Classroom Teaching Experience	11.66 (9.69)	9.74 (9.80)
	Teaching in First Grade	5.45 (6.12)	4.35 (6.00)
	Teaching in Current School	6.37 (6.88)	6.84 (7.58)

~p<.10.

Student demographic data by state are summarized in Table 2. Overall, 50.6% of the students were male and 23.83% percent were language minority students. We defined language minority students as those whose primary home language is not English. Most of these students were classified as limited English Proficient, but as definitions varied from state to state we chose to use the more inclusive term and provide descriptive data on student's scores in Table 2.

Table 2

Student Demographic Data

	Sample		Gender (male)		Language Minority Students[a]									
	TSG	Control	TSG	Control	TSG					Control				
					English Proficiency Levels					English Proficiency Levels				
					1	2	3	4	5	1	2	3	4	5
Total	273	302	135	156	0	8	27	18	6	6	12	35	21	4
CA	159	202	77	101	0	7	21	15	6	1	7	32	21	4
PA	80	72	42	40	0	0	0	0	0	4	0	1	0	0
VA	34	28	16	15	0	1	6	3	0	1	5	2	0	0

[a]The California English Language Development Test (CELDT) was administered in CA and VA. The Stanford English Language Proficiency Test (SELP) was administered in PA. English Proficiency levels for CELDT: 1 = Beginning, 2 = Early Intermediate, 3 = Intermediate, 4 = Early Advanced, 5 = Advanced. English Proficiency levels for SELP: 1 = Pre-Emergent, 2 = Emergent, 3 = Basic, 4 = Intermediate, 5 = Proficient

TSG Facilitators. There were 5 TSG facilitators: 2 in CA, 2 in PA, and 1 in VA. See Table 3 for a description of the TSG facilitators. The TSG facilitators had a strong background in scientific reading research. Four had doctoral level degrees in special education or literacy and experience with reading research. One TSG facilitator had extensive district administration experience and a background in reading instruction. At the start of the project, all facilitators met to plan and organize the TSG agendas. During study implementation, conference calls were scheduled to debrief and discuss content.

Table3

TSG Facilitator Background

TSG Facilitators	Facilitated in school districts in	Education	Background	School Experience
Facilitator 1	CA	Doctorate	Reading	Teacher; Administrator
Facilitator 2	CA	Doctorate	Reading	Teacher
Facilitator 3	VA	Doctorate	Reading	Teacher
Facilitator 4	PA	Doctorate	Reading	Teacher
Facilitator 5	PA	Masters	Reading	Teacher; Administrator

Design

Randomized field trials were used to examine the impact of the TSG intervention. In Year 1 (2004-2005) the study was conducted in a school district in CA only. In year 2 (2005-2006) the study was replicated in school districts in CA, PA, and VA. Participating schools from each district (for both Years 1 and 2) were randomly assigned to either the TSG condition or the control condition. In the CA school district and PA school district, schools were matched prior to random assignment. In CA, 10 schools (6 schools in Year 1 and 4 schools in Year 2) were matched on API (Annual Performance Index) scores, ethnic composition (percentage Hispanic), and achievement scores. In the PA school district, 6 schools were matched on free/reduced lunch status and reading proficiency on the 3rd grade statewide assessment test (Pennsylvania System of Student Assessment). Schools in VA school district were not matched due to feasibility constraints. Sample in the VA school district included three schools. Two of these were small-sized schools, which were combined into one set, and the set was treated as one school for purposes of random assignment. Teachers and schools were remunerated for their participation.

TSG and Control Conditions

Reading First mandates that all teachers in Reading First schools allocate certain time for professional development efforts in reading. While Reading First required teachers in Reading First schools to receive professional development in scientifically based reading approaches, the states and districts had wide latitude in how they operationalized and implemented these professional development activities in reading. Teachers from all three school districts from the three states attended a summer institute in reading and met during the year for the contracted professional development efforts on reading, that were mandated under Reading First. In the school districts in PA and CA, participation in TSG was counted towards the required professional development hours. In VA, it was as add-on.

Another constant was the reading curriculum used in TSG and control classroom within the same school district. Open Court was used in the CA school district, Harcourt Brace in PA school district, and the Wright Group's Guided Reading program was followed in the VA school district. Guided Reading includes a teacher's use of small group and individual instruction to help students learn comprehension strategies. Students are grouped by reading ability and use leveled reading materials ("leveled texts") selected by the teacher. Guided reading lessons are typically 15 to 20 minutes in duration and divided into three main components – pre-reading, reading, and post-reading.

At the completion of the study, teachers in both TSG and control conditions responded to a survey that sought information on reading related professional development activities attended during the school year. These data are summarized in Table 4. 77% of the teachers from TSG and 57% of the teachers from control had professional development activities in comprehension. Professional development activities in vocabulary were attended by 29% of the teachers from TSG and 23% from control. Chi-square tests revealed marginally significant differences between groups on comprehension strategies, vocabulary instruction, and data driven instruction; and significant differences on intervention strategies and structured English Emersion Techniques.

Table 4

Professional Development Activities of TSG and Control Teachers

Attended professional Development Activities in	TSG (N = 39)	Control (N = 42)
Comprehension Strategies	30~	24
Vocabulary Instruction	29~	23
Phonemic Awareness	19	25
Decoding & Phonics	18	24
Fluency	24	28
Differentiating Instruction	23	24
Lesson Study-Phonemic Awareness	20	17
Lesson Study- Decoding & Phonics	24	20
Lesson Study- Fluency	24	28
Lesson Study- Comprehension Strategies	13	18
Lesson Study- Vocabulary Instruction	13	16
Intervention Strategies	20*	32
Assessment	32	32
Data Driven Instruction	30~	25
Structured English Emersion Techniques & Strategies	24**	10
Purposeful Independent Work Time Activities	22	30

~p<.10; *p<.05; **p<.001

Control Condition. Teachers in the control condition participated in scheduled school and district professional development activities. During the study, control teachers did not participate in our TSG sessions or have access to the materials. After the studies in Years 1 and 2 were completed, facilitators offered to implement TSG sessions in control schools if principals and teachers were interested. After the Year 2 study ended, TSG facilitators helped implement TSG sessions in control schools in Pennsylvania and Virginia, as they had expressed interest in implementing the TSG in their school districts.

TSG Intervention. The TSG intervention was comprised of 16 interactive sessions held at the school site twice a month from October to mid-June, for a total of 16 sessions. The first eight sessions focused on vocabulary instruction. The remainder of the sessions addressed explicit reading comprehension instruction. Each session lasted approximately 75 minutes. Sessions were conducted at the discretion of the school principal either before or after school to maximize instructional time during the school day and not to conflict with existing reading instruction or other professional development activities. On occasion, they were conducted at a time that was convenient to the participants (e.g., weekend). Teachers were required to attend a minimum of 14 sessions to continue in the study and receive compensation.

The TSG format consisted of small group meetings (three to eight participants). Each TSG meeting was conducted in an informal style to allow for open discussion and collaboration among teachers. For example, the rooms were arranged so that the

teachers sat around a table rather than in a traditional classroom configuration. In some cases, sessions were held in the school library.

A 4-step recursive process (described below) was instituted during each TSG session. This 4-step recursive process provided a common format for the TSG sessions across facilitators and sites, while leaving room for flexibility to respond to issues or concerns specific to the site or individual teacher. For example, in one school, there were a significant number of students who spoke Spanish as their first language. The teachers had a discussion about whether the students' limited English would affect their understanding of specific words and whether they could explain the words in the students' primary language before explaining it in English.

The 4-step recursive process entails: (1) *Debrief Previous Application of the Research,* (2) *Walk Through the Research*, (3) *Walk Through the Lesson, and* (4) *Collaborative Planning*, In the first segment, *Debrief Previous Application of the Research*, the teachers reported on their implementation of the lesson they planned collaboratively during the previous TSG session. For example, in the first vocabulary session, teachers learned how to write "student friendly" definitions (Beck, McKeown & Kucan, 2002). During the debrief portion of the second vocabulary session, the facilitator reminded the participants that the purpose of the previous session was to write student friendly definitions which they would teach to their students. The facilitator asked questions to prompt participants to share what went well, what did not work well, and how students responded to the instruction.

The purpose of the second segment, *Walk Through the Research*, was to discuss the critical instructional concepts from the reading assigned at the end of the previous session. If the teachers did not readily discuss the selection or did not address the most important and relevant aspects of the material, the facilitators were prepared to prompt the participants with specific questions geared towards discussing these issues. For example, during the second vocabulary session, participants were to be prepared to discuss the chapter in *Bringing Words to Life* (Beck et al., 2002) that addressed how to choose words they should teach. A critical issue in selecting words is the level of the word's utility in language. Beck and her colleagues have defined this concept by classifying words into one of three tiers, based on their usefulness. In an effort to direct the discussion to the concept of tiered words, the facilitator asked participants to describe the attributes of Tier 1, 2, and 3 words.

During segment three, *Walk Through the Lesson*, the participants reviewed a lesson from the core reading program's Teacher's Guide that they would be teaching before the next TSG session. Their charge was to determine how the lesson did or did not exemplify the tenets of the research they discussed in the previous segment of the session. As a group they discussed the strengths and weaknesses of the publisher's suggested lesson and how it could be modified to reflect the research. Following through with the example from the second segment, the teachers worked in pairs or triads and reviewed the lesson plan for an upcoming story to determine the utility of the words the publisher recommended be pre-taught. They were also asked to be on the

lookout for words that may need to be added to enhance their students' comprehension of the selection.

In segment four, *Collaborative Planning*, teachers worked as a whole group or in pairs to plan a lesson that incorporated the targeted research principle. During the planning portion of vocabulary session two, teachers called out words that would most likely be unfamiliar to their students. The facilitator wrote the words on a white board or chart paper. As a group, they labeled each word as either Tier 1, 2, or 3. Focusing on the Tier 2 words, the participants first determined those words that were conceptually central for comprehending the story. Using this corpus of words, they decided those that could be taught briefly and those that needed more explicit instruction. Cumulative review was incorporated into this session by assigning two words to each pair and having them develop student a student friendly definition for each word. These definitions were collected by the facilitator and copied and distributed to the participants. Their assignment for this vocabulary session was to teach their students the words using the student friendly definitions and to read the next chapter in *Bringing Words to Life: Robust Vocabulary Instruction* (Beck et al., 2002)

The content covered in these 16 TSG sessions in comprehension and vocabulary is delineated in Table 5. Though a recursive process was applied to each session (i.e., consistent use of the four session segments), content was designed to build cumulatively over the series of TSG sessions. For example, at the conclusion of the vocabulary sessions, teachers engaged in a comprehensive planning activity that

required them to apply all of the instructional concepts discussed during prior TSG sessions.

Each TSG participant received a copy of *Bringing Words to Life: Robust Vocabulary Instruction* (Beck et al., 2002), an instructional rubric for evaluating comprehension lessons, and a notebook with selected research-based applied readings in vocabulary and comprehension (listed in Figure 1). Children's literature trade books were also provided to teachers for a vocabulary lesson planning activity. The children's books were selected from a "recommended" list in *Brining Words to Life*. Paper and a pen were also provided for participants to take notes during the sessions.

The goal of the TSG was to help teachers begin to think about and ultimately to use research-based instructional strategies in their classrooms by integrating the TSG content into their existing curriculum. Therefore, the purpose of the TSG was not to change a district's core curriculum, but to *enhance* implementation of that curriculum (Gersten & Woodward, 1990) by using research based strategies that may not be included in the teacher's guide to the reading series.

Table 5

Teacher Study Group Content

Session	Content	Specific Topics
Introductory Session		• Introductions • Project Overview • Teacher Surveys
1	Vocabulary	• Student Friendly Definitions • Examples and Contrasting Examples
2	Vocabulary	• Choosing Words to Teach • Tier 1, 2, and 3 Words
3	Vocabulary	• Activities that Promote Interaction with Target Words
4	Vocabulary	• Putting It Together Part 1: Planning a Trade Book Vocabulary Lesson-Generating Student Friendly Definitions, Examples and Contrasting Examples
5	Vocabulary	• Putting It Together Part 2: Planning a Trade Book Vocabulary Lessons-Generating Activities that Promote Interaction with Target Words
6	Vocabulary	• Deriving Meanings from Context
7	Vocabulary	• Creating Rich Vocabulary Environments • Putting it All Together: Planning a Core Reading Program Vocabulary Lessons- Generating Student Friendly Definitions, Examples and Contrasting Examples, Generating Activities that Promote Interaction with Target Words
8	Vocabulary	• Teaching Vocabulary to English Language Learners • Putting it All Together: Planning a Core Reading Program Vocabulary Lessons- Generating Student Friendly Definitions, Examples and Contrasting Examples, Generating Activities that Promote Interaction with Target Words
9	Comprehension	• Explicit Comprehension Instruction
10	Comprehension	• Explicit Comprehension Instruction
11	Comprehension	• Asking Questions
12	Comprehension	• Main Idea
13	Comprehension	• Question Answer Relationships (QAR)
14	Comprehension	• Question Answer Relationships (QAR)
15	Comprehension	• Generating and Evaluating Predictions: Direct Reading and Thinking Activity (DRTA)
16	Comprehension	• Pulling it All Together • The Importance of Instructional “Scaffolding”: An Example

Figure 1

Readings for Vocabulary and Comprehension

Vocabulary

1. Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction.* New York: Guildford Press.
2. Gersten, R., & Geva, E. (April, 2003). Teaching Reading to Early Language Learners, *Educational Leadership,* 44-49.

Comprehension

1. Activities from the *First Grade Teacher Reading Academy*, Vaughn Gross Center for Reading/Language Arts, University of Texas, Austin College of Education, 2002.
2. Raphael, T. E. (1986). Teaching question answer relationships, revisited. *The Reading Teacher,* 516-522.
3. Duffy, G. G. (2002). The case for direct explanation of strategies. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices (pp. 28-41).*
4. Pardo, L. (2004). What every teacher needs to know about comprehension. *The Reading Teacher, 58,* 272-280.
5. Vaughn, S. & Linan-Thompson, S. (2004). *Research-Based Methods Of Reading Instruction: Grades K-3.* Association for Supervision and Curriculum Development.
6. Barton, J., & Sawyer, D. M. (2003). Our students *are* ready for this: Comprehension instruction in the elementary school. *The Reading Teacher, 57,* 334-347.

Training of Data Collectors and Classroom Observers

All data collectors and classroom observers had classroom teaching or school experience. Data collectors were trained in conducting student assessments in mid-September at all three sites. During a 5-hour training, all data collectors were taught to administer the student measures. Procedures were standardized for all data collectors. The rules for administration and scoring of the student measures were taught, followed by demonstrations of how to administer and score the tests. Participants practiced administering and scoring each subtest at least twice. At the end of training, all potential testers were observed giving the tests in a mock session. Participants were trained until 100% accuracy was obtained on a training checklist.

Classroom Observers participated in a two-day training session. Training began with a general description of the *Reading Comprehension and Vocabulary (RCV) Observational Measure* (Gersten, Dimino, & Jayanthi, 2007). Participants were first trained in the comprehension scale and then on the vocabulary scale. Training included a discussion of the major constructs of effective reading instruction (i.e., explicit instruction, student practice, and corrective feedback), explanation of the items in the measure, a brief review of the reading comprehension and vocabulary strategies that were addressed in the measure, and rules for coding the comprehension and vocabulary items. Each participant received a codebook explaining the measure and the coding rules.

To lessen observers' anxiety, coding practice was scaffolded to ease them into the process. For example, observers' first viewed some short segments of instruction

chosen from unrehearsed reading instruction footage filmed in first grade classrooms (2-3 minutes). Then they coded longer teaching segments: first by coding some teaching segments during the second viewing, and later additional ones during the first viewing.

Coding answer keys were developed for each of these segments. These keys consisted of the time the coding occurred and the exact teacher language that generated the code. After coding, the trainer replayed the teaching segments, discussed the rationale for the codes, answered questions, and addressed concerns. At the end of training, two reliability checks, which involved coding 30 minute teaching segments, were conducted to ensure observer competency with the observation measure.

Data Collection

Student assessments were administered over a three-week period in Fall and Spring of Years 1 and 2. All measures were administered individually to the randomly selected students from each class. Testing was done outside of class in a quiet room. On the first day of testing, a testing coordinator conducted individual reliability checks with each data collector and checked each protocol to insure proper scoring. All student protocols were “double checked” for accuracy by a research member of the team, shortly after administration.

Classroom observations were conducted in each classroom (n=81) during April and early May of Years 1 and 2. All teachers were observed once; 30% of all the teachers were observed twice, and 1/8th of the teachers were observed by two observers to collect data for inter-observer reliability. Inter-observer reliability was on

average 84.49% for the vocabulary scale and 90.89% for the comprehension scale. Quality control checks for the observations were conducted during the first ten days of observations by the members of the research team to ensure desired level of accuracy and to correct possible errors in coding. Feedback was provided to the observer immediately following the observation.

All TSG sessions were tape-recorded. Tapes from sessions that were identified as having relevant and illuminating qualitative data were transcribed. The audiotapes were also used to document fidelity of implementation. At the conclusion of each session, the TSG facilitator added a reflection to the end of the audiotape. The reflection addressed questions such as, “How did the session go today?” “Was there anything unusual about this session?” “What would you change if you conducted this session again?”.

Implementation Fidelity

To determine implementation fidelity, our research staff listened to one-fourth of the audio taped TSG sessions. The same 4 TSG sessions (2 vocabulary TSG sessions and 2 comprehension TSG sessions) were reviewed for each of the nine sites. We purposely chose 4 lessons that facilitators believed would clearly demonstrate each of the key components of the TSG sessions. We specifically did not choose lessons that would be difficult to assess due to more complex content. Of the 36 tapes chosen for the fidelity check, 33 were available for review due to missing data or audio taping malfunction. On average, 86.5% of the key components were fully implemented. The fidelity means for each TSG session ranged from 83.3% to 93.8%. (See Appendix B.)

Our review of implementation reinforced our sense that fidelity was more difficult to maintain at some site than others. For example, in some schools it was difficult to schedule a full 75-minute TSG session due to school scheduling constraints and district policies about teacher release time. In some of our schools, TSG sessions were only implemented during 30-minute planning times. Under these circumstances, TSG sessions had to be continued across multiple planning days. Not surprisingly, fidelity scores were lowest in districts where the TSG facilitators were limited to 30 minutes per session with teachers. Limited sessions made implementing the full lesson difficult. Facilitators needed more time to cover and apply the material.

Additionally, although all sites were in Reading First districts, sites with lower fidelity scores were in districts that did not use a core reading series. A core reading series enabled facilitators to implement all of the key components of the TSG sessions. For example, one of the key components of the sessions is the time for teachers to plan an upcoming lesson collaboratively. If the teachers were using a core reading series, they could collaboratively plan a lesson that was from an up-and-coming unit. In districts where no core reading series was required, teachers did not have similar future lessons to discuss. Implementing collaborative planning was difficult as teachers didn't follow a sequence of prescribed lessons or couldn't identify specific lesson content. Overall, we found that sites with the highest fidelity scores were in districts where the TSG facilitators were given 75 minutes to meet with teachers after school and where the district mandated the use of a core reading series.

Measures

Teacher Measures. We used the *Reading Comprehension and Vocabulary (RCV)* Observational Measure (Gersten et al., 2007) as a posttest of teaching practice in comprehension and vocabulary. The 34-item comprehension scale has an internal consistency coefficient of .69. The 12-item vocabulary scale has an internal consistency coefficient of .70. We developed the *RCV* measure, a moderate-inference measure, to assess the quality of classroom reading instruction. The items in the measure 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). The measure is well aligned with the extant literature on effective reading instruction (e.g., Anderson, Evertson, & Brophy, 1979; Baumann & Kameenui, 1991; Beck et al., 2002; Graves, 2006).

We used the *Content Knowledge for Teaching Reading (CKTR)* assessment (Phelps & Schilling, 2004) as a posttest to measure teacher knowledge in vocabulary and comprehension. The *CKTR*, has alpha reliability coefficients in the range of .67 to .82; The IRT estimated reliabilities are above .70 for each scale. Teachers are provided classroom scenarios or instructional examples and asked questions that relate to instructional decisions based on research-supported practices.

We utilized two scales from the surveys developed by the Consortium on Chicago School Research (2000) to examine the impact of the TSG on teacher perceptions of professional culture. The two scales include, the *Quality Professional*

Development scale and the *Teacher-Teacher Trust* scale. The *Quality Professional Development* scale measures teachers' perceptions of the extent to which professional development has influenced their teaching and understanding of their students, and provided them with opportunities to work with their colleagues. The *Quality Professional Development* scale has 9 items and an internal consistency coefficient of .93[2]. The *Teacher-Teacher Trust* scale measures the degree to which teachers care and have mutual respect for each other, and the extent to which they are comfortable in sharing their concerns with each other. It has 6 items and an internal consistency coefficient of .90[3]. We modified the wording in these 6 items to reflect the grade level interactions that were central to the TSG intervention. For example, "Teachers in this school trust each other" was changed to "Teachers in this grade level trust each other". See Figure 2 for a sample listing of these items. All items were likert scale type items.

[2]Scale reliability reported by The Consortium on Chicago School Research (2000) = .84.

[3]Scale reliability reported by The Consortium on Chicago School Research (2000) = .82.

Figure 2

Teacher Perceptions of Professional Culture: A Sample Listing of items

Scale: Quality Professional Development

1. Overall, my professional development experiences over the past year have included opportunities to work productively with teachers from other schools.

Strongly Disagree Disagree Agree Strongly Agree

2. Overall, my professional development experiences have included enough time to think carefully about, to try, and to evaluate new ideas.

Strongly Disagree Disagree Agree Strongly Agree

3. Most of what I learn in professional development addresses the needs of the students in my classroom.

Strongly Disagree Disagree Agree Strongly Agree

4. Overall, my professional development experiences have deepened my understanding of subject matter.

Strongly Disagree Disagree Agree Strongly Agree

Scale: Teacher-Teacher Trust

1. Teachers in this grade level trust each other.

Strongly Disagree Disagree Agree Strongly Agree

2. It's OK in this grade level to discuss feelings, worries, and frustrations with other teachers.

Strongly Disagree Disagree Agree Strongly Agree

3. Teachers respect other teachers who take the lead in grade level improvement efforts.

Strongly Disagree Disagree Agree Strongly Agree

4. Teachers in this grade level respect those colleagues who are expert at their craft.

Strongly Disagree Disagree Agree Strongly Agree

Source: Consortium on Chicago School Research (2000)

Student Measures. To assess early literacy skills, students were administered three Dynamic Indicators of Basic Early Literacy Skills (DIBELS) measures (Good & Kaminski, 2002; Kaminski & Good, 1996). Measures selected were *Letter Naming Fluency*, *Phonemic Segmentation Fluency,* and *Oral Reading Fluency.* These measures are routinely used to screen students for difficulties in early reading. *Letter Naming Fluency (LNF)* is a 1-minute timed measure that assesses the accuracy and speed with which children identify letter names. Examiners present randomly ordered upper- and lower-case letters and ask the child to name as many letters as possible in 1 minute. LNF is calculated as the number of correct letter names per minute. *LNF 6th Edition* has test-retest reliability of .88, and a predictive validity of .65 for reading performance a year later. Alternate form reliability reported for use of LNF with kindergarten students is .99.

Phonemic Segmentation Fluency (PSF) is a 1-minute timed measure that assesses a student's ability to segment fluently regular three-to-four phoneme words into individual phonemes. The examiner orally presents words (i.e., three-to-four phoneme words) and the child has to respond by saying the individual phonemes in each word. For example, the examiner would say "cat." To answer correctly, students would say "/k/ /a/ /t/", with scores calculated according to the number of correct segments identified per minute. *PSF* has a test-retest reliability of .88 and a predictive validity of .68 for end of first grade reading on the *Woodcock Johnson*. Alternate-form reliability for the *PSF* is reported at .88. Criterion-related validity coefficients range from .43 to .67.

Oral Reading Fluency (ORF) is a 1-minute timed measure that assesses a child's ability to read grade level passages fluently and accurately. During the ORF assessment, children are asked to read an excerpt of approximately 60 words of connected text in 1 minute. Reading performance is determined by scoring the total number of words read correctly per minute. *ORF* has a test-retest reliability in the .90s (.92 to .97); alternate form reliability of different reading passages drawn from the same level ranged from .89 to .94 (Tindal, Marston, & Deno, 1983). Criterion-related validity is from .52 to .91 (Good & Jefferson, 1998).

The following subtests of the *Woodcock Diagnostic Reading Battery (WDRB)* were also administered: *Word Attack*, *Letter-Word Identification*, *Reading Vocabulary*, *Oral Vocabulary*, and *Passage Comprehension*. Internal consistency for all these subtests is above .90.

Interviews with TSG Participants

TSG participants from the CA school district were interviewed individually at the end of the study about their experiences with the TSG and other professional development activities. Interviews were audio taped and teachers where informed that there were no "right" or "wrong" answers. Teachers were also assured that their individual interviews would not be heard by their TSG facilitators or school officials. Interview questions asked included how the TSG compared to other school or district professional development activities, which topics in the vocabulary and comprehension sessions were most useful, which TSG strategies were "tried out" in their classrooms, and which strategies the teachers thought they would continue to use in their teaching.

Teachers also discussed the challenges of committing to a series of TSG sessions and whether they would participate in the TSG again.

Results

Sample Description and Baseline Characteristics of Students and Schools

A total of 84 teachers in 10 TSG and 9 control schools were included at the beginning of the study. The final analytic sample was 81 teachers. 575 children completed individually administered pretests on the three *DIBELS* measures: *Letter Naming Fluency (LNF)*, *Phonemic Segmentation Fluency (PSF)*, and *Oral Reading Fluency (ORF)*. Descriptive statistics in Table 6 provide means and standard deviations for these three potential covariates. Scores on the *LNF* measure were normally distributed whereas scores on *ORF* displayed floor effects due to the large number of children who scored zero on the fall pretest. Validation studies of the *DIBELS* indicate that first-grade performance on the *LNF* is also a stronger predictor of achievement on standardized tests of reading comprehension (e.g., Stanford Diagnostic Reading Test, Metropolitan Reading Test) than *PSF* (Good, Gruba, & Kaminski, 2001, p. 684). Additional research on early literacy suggests that *LNF* is the best predictor of reading achievement at the end of first grade (Bond & Dykstra, 1967/1997). Therefore, we used scores from the *LNF* test as the covariate in the models to estimate treatment effects on the student reading outcomes.

We conducted t-tests to compare school means on the pretest reading measures and to assess the equivalence of the two groups at the beginning of the experiment. As shown in Table 7, there was no statistically significant difference between TSG and

control schools on either the *DIBELS* or the *WDRB* pretest measures. These results suggest that random assignment of schools to experimental conditions created two equivalent groups of schools at the beginning of the study.

Table 6

Means and Standard Deviations for Pretest Student Reading Measures

	N	Min	Max	Mean	SD
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)					
Letter Naming Fluency (LNF)	575	0	91	37.80	15.42
Phonemic Segmentation Fluency (PSF)	575	0	74	26.17	16.87
Oral Reading Fluency (ORF)	575	0	109	9.46	12.48
Woodcock Diagnostic Reading Battery (WDRB)					
Letter Word Identification (LWID)	575	316	488	403.33	35.15
Word Attack	573	343	495	449.90	17.59
Reading Vocabulary	575	425	491	437.72	13.91
Passage Comprehension	575	380	484	411.41	25.26
Oral Vocabulary	574	404	509	452.06	17.22

To assess threats to internal and external validity, we undertook two additional analyses. First, we examined whether the percentage of children with missing posttest scores (primarily due to family moves) was similar between conditions. The final analytic sample was based on a total sample size of 468 students. Although 18.6% of students (n = 107) missed posttests, the chi-square analysis revealed no significant relationship between the proportion of children with missing data in the TSG group (n = 56) and the control group (n = 51), χ^2 (1,575) = 1.24, p = .265. Note that 79.5% of the TSG group and 83.1% of the control group remained in the analysis of student outcomes.

Table 7

Characteristics of Treatment and Control Schools at the Beginning of the Study

Measure	Condition	N	Mean	SD	*t*	*df*	*p*	95% Confidence Interval	
DIBELS									
Letter Naming Fluency (LNF)	Treatment	10	36.45	5.20	-0.573	17	0.574	-7.045	4.037
	Control	9	37.96	6.24					
Phonemic Segmentation	TSG								
Fluency (PSF)		10	25.69	6.97	0.733	17	0.474	-4.782	9.874
	Control	9	23.15	8.17					
Oral Reading Fluency (ORF)	TSG	10	9.03	3.06	-0.392	17	0.700	-4.233	2.908
	Control	9	9.69	4.28					
WDRB									
	TSG		438.9						
Reading Vocabulary		10	4	7.03	0.178	17	0.861	-6.590	7.803
	Control		438.3						
		9	3	7.84					
	TSG		412.7						
Passage Comprehension		10	8	13.64	-0.036	17	0.972	-14.625	14.134
	Control		413.0						
		9	2	16.07					
	TSG		452.8						
Oral Vocabulary		10	4	7.18	-0.067	17	0.947	-8.475	7.950
	Control		453.1						
		9	1	9.72					

Second, we compared the pretest scores of students who remained in the study and those who moved during the school year. There was no statistically significant difference between students who remained in the study and those who were excluded from the final analysis on the measure of *Letter Naming Fluency (LNF)* ($t(573) = -1.29$, $p = .197$), *Phonemic Segmentation Fluency (PSF)* ($t(573) = .132$, $p = .895$), and Oral

Reading Fluency *(ORF)* (*t*(573) = -.12, *p* = .905). Similarly, there was no significant difference between groups on the *WDRB* measures of reading vocabulary and passage comprehension.

Treatment Effects on Teacher and Student Outcomes

Recently, multilevel models have been widely employed in cluster-randomized field trials to estimate the efficacy of school-level interventions on teacher practice and student achievement Bloom, Richburg-Hayes, & Black, 2007; Borman et al., 2005). Since our study of the TSG involved the random assignment of schools to TSG and control conditions, we employed a two-level model to estimate treatment effects on teacher and student outcomes. Since students and teachers within a school share common experiences, their outcomes are likely to be correlated. Thus, our multi-level models included individual- and group-level error terms to account for the clustering of teachers within schools and students within schools. The inclusion of both an individual and group-level error term is needed to estimate correct standard errors used in hypothesis tests of whether the difference between TSG and control schools was significantly different from zero. 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.

Impact Estimates on Teacher Outcomes. We employed a two-level model to estimate the impact of the Teacher Student Group on measures of teacher practice and

teacher knowledge. The fully unconditional model at Level 1 for teacher *i* in school *j* can be written as

$$Y_{ij} = b_{oj} + \varepsilon_{ij}, \quad (1),$$

and the Level 2 model for the intercept is

$$ß_{oj} = g_{00} + g_{01} (TSG)_j + \mu_{0j}, \quad (2)$$

where $ß_{oj}$, the mean score on the teacher outcome for school *j*, is regressed on the dummy variable, Teacher Study Group (TSG), which takes on a value of one for schools assigned to the experimental condition and a value of zero for the control condition. Our goal here is to estimate the treatment effect, which is captured by the level 2 parameter, g_{01}. The Level 1 and Level 2 can be combined to form the following mixed-effects model

$$Y_{ij} = g_{00} + g_{01}(TSG)_j + (\mu_{0j} + \varepsilon_{ij}), \quad (3)$$

where μ_{0j} is a random effect for school *j* and ε_{ij} is the teacher-specific error term for teacher *i* in school *j*. The treatment dummy variable is modeled as a fixed effect and the teacher and school residual terms are modeled as random effects. Using the third equation, we estimated the treatment effect on the measures of teacher practice and teacher knowledge.

Impact on Observed Teaching Practice. Table 8 displays the magnitude of the estimated treatment effect on the reading comprehension and vocabulary observation measures used to assess the quality of instruction in TSG and control schools. The coefficient for the treatment dummy variable indicates that teachers in TSG schools scored .86 standard deviations higher on the comprehension measure and .58 standard

deviations higher on the vocabulary measure relative to teachers in control group schools.[4]

Table 8

Estimated Treatment Effects on Observed Teaching Practice (Comprehension and Vocabulary)

	Reading Comprehension			Vocabulary		
Measures	Coefficient	se	t ratio	Coefficient	se	t ratio
Fixed Effect						
Intercept, g_{00}	-0.40	0.18	-2.27*	-0.28	0.15	-1.90
Teacher Study Group, g_{01}	0.86	0.25	3.43**	0.58	0.21	2.74**
Random Effect	Variance Component	c*2*	*df*	Variance Component	c*2*	*df*
Between-school variance, m_{0i}	0.13	2.53~	17	<1.00	0.00	17
Within-school variance, e_{ij}	0.64			0.90		

~p<.10, *p<.05, **p<.01

Impact on Teacher Knowledge of Reading Instruction. Results from the multilevel models used to estimate the treatment effect on the teacher knowledge measures of comprehension and vocabulary instruction are presented in Table 9. Although the effect size of .32 suggests that teachers in the TSG schools scored higher on the measure of comprehension knowledge, this standardized mean difference was not significantly different from zero. However, the effect on knowledge of vocabulary

[4]We replicated the analysis using MANOVA and obtained results that were similar to those in Table 8.

instruction was significant. Teachers in the TSG schools outperformed teachers in the control schools by approximately .73 standard deviations on the teacher knowledge measure of vocabulary instruction. As with the teacher observation measures, there was no significant variability across schools on the teacher knowledge measures for comprehension and vocabulary.[5]

Table 9

Estimated Treatment Effects on Teacher Knowledge (Comprehension and Vocabulary)

	Reading Comprehension			Reading Vocabulary		
Measures	Coefficient	se	t ratio	Coefficient	se	t ratio
Fixed Effect						
Intercept, g_{00}	-0.19	0.20	-0.93	-0.42	0.22	-1.93~
Teacher Study Group, g_{01}	0.32	0.28	1.11	0.73	0.30	2.42*
Random Effect	Variance Component	c*2*	*Df*	Variance Component	c*2*	*df*
Between-school variance, m_{0i}	0.12	1.48	17	0.23	2.56~	17
Within-school variance, e_{ij}	0.86			0.72		

~p<.10, *p<.05, **p<.01

***Impact of the TSG on Teacher Perceptions of Professional Culture*.** Results from the multilevel models used to estimate the treatment effect on the *Quality*

[5]We replicated the analysis using MANOVA and obtained results that were similar to those in Table 9.

Professional Development scale and *Teacher-Teacher Trust* scale are presented in Table 10. Each of the multilevel models also includes a pretest score on the survey measure, which improved the precision of the estimated treatment effect. Our findings suggest that teachers in the experimental condition expressed significantly more positive views toward professional development (*ES* = .39) than teachers in the control condition. Since there was significant between school variability on this measure, the multilevel model was the appropriate analytic strategy for estimating the impact on teachers' overall views toward professional development. However, there was no significant difference between groups on the scale measuring teachers' trust and respect for each other.

Table 10

Estimated Treatment Effects on Teacher Perceptions of Professional Culture

	Quality Professional Development			Teacher-Teacher Trust		
Measures	Coefficient	se	t ratio	Coefficient	se	t ratio
Fixed Effect						
Intercept, g_{00}	-.01	.16	-.07	-.12	.22	-.55
Pretest Score, g_{01}	.45	.07	6.06**	.30	.11	2.76*
Teacher Study Group, g_{02}	.39	.22	1.76~	.20	.30	.65
Random Effect	Variance Component	c*2*	*df*	Variance Component	c*2*	*df*
Between-school variance, m_{0j}	.16	7.65	17	.35	4.71	17
Within-school variance, e_{ij}	.28			.53		

~p<.10, *p<.05, **p<.01

Impact Estimates on Student Outcomes. We specified a two level model to estimate treatment impacts on the student outcome measures of reading. Formally, the Level 1 model for student *i* in school *j* can be written as

$$Y_{ij} = b_{oj} + \varepsilon_{ij}, \quad (1)$$

where Y_{ij} is the posttest reading score for student *i* in school *j*, b_{oj} is the mean posttest score for school *j*, and ε_{ij} is the error term for student *i* in school *j*. The fully specified Level 2 model is written as

$$ß_{oj} = g_{00} + g_{01}(LNF)_j + g_{02}(TSG)_j + \mu_{0j}, \quad (2)$$

where, $ß_{oj}$ is the posttest reading score for school j and predicted by a pretest covariate, the school mean scores on *Letter Naming Fluency (LNF)* measures, and the treatment dummy variable denoting whether a school was randomly assigned to a control or experimental condition. Inclusion of the pretest score improved the precision of the estimated treatment effect, which is captured by the level 2 parameter g_{02}. To create an unbiased impact estimate, Level 1 and Level 2 can be combined to form a mixed-effects model, which can be written as

$$Y_{ij} = g_{00} + g_{01}(LNF)_j + g_{02}(TSG)_j + (\mu_{0j} + \varepsilon_{ij}), \quad (3)$$

where the pretest *LNF* score and the treatment dummy variable are modeled as fixed effects and the student and school residual terms are modeled as random effects.

We used the XTREG routine for multilevel modeling in Stata to conduct the impact analyses on student outcomes. We estimated the impact of the Teacher Study Group intervention on posttest measures of reading achievement from the *WDRB* and *DIBELS*. The battery included three measures that were not directly related to the intervention focus: *WDRB Letter Word Identification, WDRB Word Attack,* and *DIBELS Oral Reading Fluency*, and another three *WDRB* measures that were directly related to the focus of the intervention: *Reading Vocabulary, Oral Vocabulary*, and *Passage*

Comprehension. As noted earlier, we standardized all measures to have a mean of one and standard deviation of zero. Thus, the coefficient for the treatment dummy variable can be interpreted as an effect size, that is, the standardized mean difference in the relevant posttest outcome between treatment and control schools.

As shown in Table 11, there was no statistically significant impact on measures of *Letter-Word Identification, Word Attack,* and *Oral Reading Fluency*, that is, the non-target outcomes. However, the magnitude of the treatment effects, which ranged from .13 to .22, fall in line with impact estimates from recent cluster-randomized trials of school-level interventions like Success for All (Borman et al., 2005).

Table 12 displays results for the multilevel model used to estimate treatment effects on *Reading Vocabulary, Oral Vocabulary*, and *Passage Comprehension*. The results revealed no significant impact on the posttest *WDRB* measures of *Reading Vocabulary* and *Passage Comprehension*. However, the moderately large effect size for *Oral Vocabulary*, ES = .44, was marginally significant.

We also conducted a correlational analysis to assess the relationship between the teacher measures and the average performance level of children in a classroom. Table 13, displays correlations between each of the two teacher measures and classroom means for the student reading outcomes, partialling out initial pretest scores on the *Letter Naming Fluency (LNF)* assessment. We refer to these as adjusted growth scores. We found several significant, moderately sized correlations between the teacher measures and student reading outcomes. Scores on the teacher knowledge measure of comprehension and vocabulary were significantly associated with mean classroom

Table 11

Estimated Treatment Effects on Reading Accuracy and Fluency

	Letter Word Identification			Word Attack			Reading Fluency		
Measures	Coefficient	se	t ratio	Coefficient	se	t ratio	Coefficient	se	t ratio
Fixed Effect									
Intercept, g_{00}	-0.18	0.14	-1.29	-0.14	0.16	-0.88	-0.16	0.11	-1.45
Mean Letter Naming Fluency, g_{01}	0.49	0.04	12.25	0.38	0.04	9.50	0.52	0.04	13.00
Teacher Study Group, g_{02}	0.21	0.19	1.11	0.13	0.22	0.59	0.22	0.15	1.47
Random Effect	Variance Component	c*2*	*df*	Variance Component	c*2*	*df*	Variance Component	c*2*	*df*
Between-school variance, $m_{\square i \square}$	0.14	63.75	16	0.20	85.15	16	0.08	30.88	16
Within-school variance, e_{ij}	0.62			0.62			0.59		

~p<.10, *p<.05,
**p<.01

adjusted growth scores on all *WDRB* measures and *Oral Reading Fluency*. Scores on the teacher observation scale for both comprehension and vocabulary were significantly correlated with mean classroom adjusted growth scores on *Oral Reading Fluency, Letter Word Identification, Word Attack*, and *Reading Vocabulary*. Scores on the teacher observation scale for vocabulary instruction were also correlated with *Passage Comprehension* scores.

Table 12

Estimated Treatment Effects on Vocabulary and Passage Comprehension

	Reading Vocabulary			Oral Vocabulary			Passage Comprehension		
Measures	Coefficient	se	t ratio	Coefficient	se	t ratio	Coefficient	se	t ratio
Fixed Effect									
Intercept, γ_{00}	-0.15	0.14	-1.07	-0.22	0.18	-1.22	-0.12	0.13	-0.92
Mean Letter Naming Fluency, γ_{01}	0.42	0.04	10.50	0.29	0.04	7.25	0.46	0.04	11.50
Teacher Study Group, γ_{02}	0.21	0.20	1.05	0.44	0.25	1.76~	0.12	0.19	0.63
Random Effect	Variance Component	χ^2	*df*	Variance Component	χ^2	*df*	Variance Component	χ^2	*df*
Between-school variance, $\tau_{\beta\beta}$	0.15	57.21	16	0.27	81.54	16	0.14	57.88	16
Within-school variance, σ^2_{ij}	0.64			0.69			0.61		

~p<.10, *p<.05, **p<.01

Table 13

Classroom-Level Correlations Between Teacher Measures and Student Reading Outcomes, Controlling for Pretest Scores on Letter Naming Fluency

	Teacher Practice		Teacher Knowledge	
Student Reading Outcomes	Comprehension	Vocabulary	Comprehension	Vocabulary
Oral Reading Fluency	0.33**	0.35**	0.23*	0.29*
Letter Word Identification	0.26*	0.29*	0.34**	0.31**
Word Attack	0.31**	0.32**	0.36**	0.28*
Reading Vocabulary	0.24*	0.23~	0.22~	0.27*
Passage Comprehension	0.08	0.27*	0.34**	0.31**
Oral Vocabulary	0.21~	0.20	0.41**	0.49**

~p < .10, *p < .05, **p < .01

Professional Appraisal of TSG

Overall, participants felt positive about their participation in the TSG. Most felt that the TSG was much more useful and beneficial than other forms of professional activities they experienced. Majority of the participants noted that they would volunteer for a TSG, if one were to be held at their school on another topic. Participant responses are summarized in Table 14.

Table 14

Professional Appraisal of TSG

Question	Percentage* of TSG Participants			
Helpfulness of the TSG in teaching reading	Not at all helpful	2	4	Very helpful
	0	0	26	72
Most helpful features of TSG	Least Helpful	3rd Most Helpful	2nd Most Helpful	Most Helpful
Debrief	31	33	13	20
Walk through the research	23	38	18	18
Walk through the lesson	33	15	20	28
Collaborative Planning	5	13	46	33
Volunteer for a future TSG	Definitely Not Volunteer	Might Volunteer	Probably Volunteer	Definitely Volunteer
	5	5	20	64
Implement TSG skills in the future	Sometimes	Most of the time		All of the time
Vocabulary	8	64		26
Comprehension	10	59		28
When compared to other professional activities, TSG is	Somewhat Beneficial	Mostly Beneficial		More Beneficial
	3	26		97

**Percentages have been rounded off to the nearest whole number and do not total to 100 due to missing data.*

Summary and Conclusions

We concur with Wayne and colleagues (Wayne, Yoon, Zhu, Cronen, & Garet, 2008) that despite the complexities entailed in both design and implementation, large scale randomized controlled trials are critical in the field of professional development to assess whether professional development programs have the intended impacts on classroom teaching and student achievement. The purpose of this study was to rigorously evaluate the impact of the Teacher Study Group (TSG), a professional

development intervention, on the vocabulary and reading comprehension instruction of first grade teachers in Reading First classrooms.

We intentionally chose to focus on vocabulary and reading comprehension because, when the study began, most professional development efforts for first grade teachers were focused on phonemic awareness, decoding, and strategies for building reading fluency. Additionally, reading comprehension and vocabulary are excellent topics for dynamic study group discussions and activities. We also noted that researchers had found that effecting change in comprehension and vocabulary instruction was particularly difficult (Gersten, Vaughn, Deshler, & Schiller, 1997; Klinger, Vaughn, & Hughes, 1999; Carlisle & Rice, 2002).

For this study, we assessed the impact of the Teacher Study Group on teachers' knowledge of effective vocabulary and comprehension instruction using an instrument developed by Phelps and Schilling (2004). Most importantly, we assessed the impact on observed teaching practice in the areas of comprehension and vocabulary using a measure developed by our research team (Gersten et al., 2007). Finally, we assessed impacts on student reading achievement with a particular focus on the areas of comprehension and vocabulary. We also included measures of other critical areas of reading instruction in first grade such as decoding and oral reading fluency.

Results indicated significant impacts on both classroom observation scales with effect sizes of .86 for comprehension and .58 for vocabulary. Both subscales demonstrated reasonable reliability (.69 for comprehension and .72 for vocabulary), suggesting that they measure coherent constructs. Data indicate that teachers were

implementing at least some of the types of interactive explicit instruction that were promoted in the TSG.

For the teacher knowledge measure, only the impact on knowledge of vocabulary instruction was significant (effect size of .73). However, the impact on comprehension was in the expected direction, .32. We believe the critical factor that led to significant effects in vocabulary knowledge was the cumulative review of the vocabulary concepts from one book, *Bringing Words to Life* (Beck et al., 2002). During each vocabulary session, teachers developed and practiced the research concepts that were discussed in previous study group sessions. For example, the first session addressed developing student friendly definitions, examples and contrasting examples. The second session focused on choosing words to teach before students read a selection, but also provided teachers with an opportunity to develop student friendly definitions, examples and non-examples for the target words. The same procedure was implemented during the third session, where the focus was activities to promote interaction with words. In this session, the participants chose words and then developed student friendly definitions, examples, contrasting examples, and activities to promote interaction with the words they choose. This iterative procedure appeared to foster automaticity in planning and executing teaching behaviors associated with effective vocabulary instruction (Beck et al).

Because we could locate no comparable book in the area of comprehension, we needed to rely on a series of articles. The set of articles did not– and probably could not– provide the type of coherence that the Beck et al volume did. We were able to

locate several excellent books on comprehension instruction (e.g. Carlisle & Rice, 2002; Mandel, Morrow, Gambrell, & Pressley, 2003; Sweet & Snow, 2003; Pressely, 2002; Stanovich, 2000), but all of them seemed better suited for a graduate course than an ongoing professional development course. We do see a need for such a book to accompany professional development in the area of comprehension.

Due to the manner in which comprehension is taught in core reading programs, each comprehension strategy was covered in only one session. Consequently, incorporating cumulative review was not possible. The activity that was consistent across the comprehension sessions was asking participants to analyze the comprehension instruction in their core reading program by responding to a consistent set of *Guiding Questions*. (Table 14). These questions address whether or not the tenets of explicit reading comprehension instruction are present in the Teacher's Guide. During the *Collaborative Planning* portion of the sessions, participants enhanced the lessons by incorporating the instructional features they determined were lacking.

Even though participants completed the *Guiding Questions* for several reading comprehension strategies, the recursive activity did not seem to increase their knowledge of reading comprehension. The results of the comprehension portion of the knowledge measure suggest that participants need to review and practice a strategy (e.g., main idea) several times. To impact participants' knowledge, study groups need to review comprehension instruction concepts and apply them more than once.

Table 14

Guiding Questions for Comprehension Instruction

Guiding Question		What will you do to incorporate the research?
Does the lesson explicitly explain what the strategy is, when it would be used, and the steps for doing the strategy?	Yes No Somewhat	
Does the lesson include a teacher think-aloud; i.e., repeatedly stating and modeling the "secret" to doing it successfully, so students can see the mental workings involved?	Yes No Somewhat	
Does the lesson provide scaffolded practice, with students having multiple opportunities to practice, gradually moving to independent strategy use?	Yes No Somewhat	
Does the lesson focus on two purposes: reading for text content and for application of the strategy?	Yes No Somewhat	
Does the lesson close with an explicit statement about the strategy and how to implement it?	Yes No Somewhat	

In contrast to the teacher knowledge measure, the effect size for the comprehension scale of the observation measure was higher than that of vocabulary. One reason for this may be that teachers could be somewhat better at teaching vocabulary in an interactive fashion than comprehension. Anecdotal evidence from our observers lends credence to this assumption. Another reason may be an artifact of the manner in which scores were calculated on the observation measure. Since the effect sizes were based on the total score for each domain, an item-by-item analysis is needed to determine if the TSG teachers implemented more of the teaching behaviors they learned in the TSG sessions than the control teachers.

The impact of the TSG on student achievement was measured using the Woodcock Diagnostic Reading Battery. An effect size of .44 was observed in the area of

oral vocabulary; it was marginally significant at $p < .10$. This effect size was double the impact on the reading vocabulary subtest of .21. One possible explanation for this seeming contradiction is that students' vocabulary knowledge was affected by the changes in teaching practice, but the limited reading proficiency of many of the students may have inhibited performance on reading vocabulary items. Effects on passage comprehension were negligible and non significant (effect size of .12).

We also found effects of .21 and .22 in word identification and oral reading fluency (passage reading). Neither is significant but both fall in the range found by other researchers (e.g. Borman et al., 2005) in large randomized controlled trials of comprehension reading programs. In this study, it is possible that the improved comprehension and vocabulary instruction led to some carryover in students' ability to read words and passages. In contrast, effects on decoding of pseudowords (a purer phonics measure) were trivial. This reflects, we believe, the fact that increased knowledge of word meaning might influence word recognition but not knowledge of the rules for decoding phonetically, since the pseudowords have no meaning.

In summary, this study demonstrates a good deal of promise for professional development models that (1) are focused on findings from scientific research, (2) are applied to the existing curriculum in a given school, and (3) facilitate collegial interactions with members of grade level teams, such as the TSG. Clearly, larger scale, more powerful studies are needed to verify the effectiveness of this approach. Nonetheless, the significant impacts on observed teaching practice in the areas of comprehension and vocabulary suggest real promise. These findings also suggest that

professional development efforts in first grade (and by implication kindergarten) can benefit from a strong vocabulary and comprehension emphasis.

References

Anderson, L., Evertson, C., & Brophy, J. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal, 79*, 193-223.

Ball, D. L. (1990). Reflections and deflections of policy: The case of Carol Turner. *Educational Evaluation and Policy Analysis*, *12*, 247-259.

Barton, J., & Sawyer, D. M. (2003). Our students *are* ready for this: Comprehension instruction in the elementary school. *The Reading Teacher, 57,* 334-347.

Baumann, J. F., & Kameenui, 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.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). Brining 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. *Elementary School Journal*, *96*, 385-414.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Blum, H., T., Yocom, D. J., Trent, A., McLaughlin, M. (2005). Professional development: When teachers plan and deliver their own**.** *Rural Special Education Quarterly, 24*(2), 18-21.

Bond, G. R. & Dykstra, R. (1997). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly, 32*, 348-427. (Original work published 1967).

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, T. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*, 673-696.

Carlisle, J., & Rice, M. (2002). *Improving reading comprehension: Research-based principles and practices*. Baltimore, MD: York Press.

Carroll, D. (2005). Learning through interactive talk: A school-based mentor teacher study group as context for professional learning. *Teaching and Teacher Education: An International Journal of Research and Studies, 21*, 457-473.

Consortium on Chicago School Research. (2000). Public Use Date Set User's Manual June 2000. Retrieved February 3, 2004, from http://www.consortium-chicago.org/surveys/pdfs/surveymanual.pdf

Duffy, G. G. (2002). The case for direct explanation of strategies. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices (pp. 28-41).*

Fernandez, C., & Chokshi, S. (2002). A practical guide to translating lesson study for a U.S. setting. *Phi Delta Kappan*, *84*, 128-34.

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*, 915-945.

Gersten, R., & Brengelman, S. (1996). The quest to translate research into classroom practice: The current knowledge base. *Remedial and Special Education, 96*, 228-244.

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., & Geva, E. (2003). Teaching reading to early language learners. *Educational Leadership, 60*(7), 44-49.

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*, 52-66.

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997). What we know about using research findings: Implications for improving special education practice. *Journal of Learning Disabilities*, 30, 466-476.

Gersten, R., & Woodward, J. (1990). Rethinking the regular education initiative: Focus on the classroom teacher. *Remedial and Special Education*, *11*, 7-16.

Goldenberg, C., & Gallimore, R. (1991). Changing teaching takes more than a one-shot workshop. *Educational Leadership, 49*(3), 69-72.

Good, R.H., Gruba, J., & Kaminski (2001). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas

& J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 679-700). Washington, DC: National Association of School Psychologist.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on Curriculum-Based Measurement validity. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement* (pp. 61-88). New York, NY: Guilford.

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.

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

Huberman, A. M., & Miles, M. B. (1984). *Innovation up close: How school improvement works*. New York: Plenum Press.

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

Klinger, J. K., Vaughn, S., & Hughes, M. T. (1999). Sustaining research-based practices in reading: a 3-year follow-up. *Remedial and Special Education, 20*, 263-274.

Knowles, M. S., Holton, E. F., & Swanson, R. A. (2005). *The adult learner: The definitive classic in adult education and human resource development, 6th Edition.* Burlington, MA: Elsevier.

Lambert, L. (2002). A framework for shared leadership. *Educational Leadership, 59*, 37- 40.

Lewis, C., Perry, R., Hurd, J., & O'Connell, M. P. (2006). Lesson study comes of age in North America. *Phi Delta Kappan, 88*, 273-281.

Mandel, L. M., Morrow, L., Gambrell, L. B., & Pressley, M. (Eds.). (2003). *Best practices in literacy instruction*. New York: The Guildford Press.

McLaughlin, M. (1990). The rand change agent study revisited: Macro perspectives and micro realities. *Educational Research, 19*(9), 11-16.

Meyer, R. J., Brown, L., DeNino, E., Larson, K., McKenzie, M., Ridder, K., et al. (1998). *Composing a teacher study group: Learning about inquiry in primary classrooms*. Mahwah, NJ: Lawrence Erlbaum.

Murphy, C. (1992). Study groups foster school-wide learning. *Educational Leadership, 50*, 71-74.

Pardo, L. (2004). What every teacher needs to know about comprehension. *The Reading Teacher, 58,* 272-280.

Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *Elementary School Journal, 105*, 31-48.

Pressley, M. (2002). *Reading instruction that works: The case for balanced teaching.* New York: The Guildford Press.

Raphael, T. E. (1986). Teaching question answer relationships, revisited. *The Reading Teacher,* 516-522.

Saunders, W., O'Brien, G., Hasenstab, K., Marcelletti, D., Saldivar, T., & Goldenberg, C. (2001). Getting the most out of site-based professional development. In P.

Schmidt & P. Mosenthal (Eds.). *Reconceptualizing literacy in the new age of pluralism and multiculturalism* (pp. 289-320). Greenwich, CT: IAP.

Stanovich, K., E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: The Guildford Press.

Sweet, A. P., & Snow, C. E. (2003). *Rethinking reading comprehension*. New York: The Guildford Press.

Tichenor, M. S., & Heins, E. (2000). Study groups: An inquiry-based approach to improving schools. *The Clearing House, 73,* 316-319.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Rep. 109). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Vaughn, S. & Linan-Thompson, S. (2004). *Research-Based Methods Of Reading Instruction: Grades K-3.* Association for Supervision and Curriculum Development.

Watanabe, T. (2002). Learning from Japanese lesson study. *Educational Leadership, 59*(6), 36-39.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*, 469-479.

Appendix A

Teacher Demographic Data by Year and State

	N	University Training	Certification	Total number of years of Means (Standard Deviations)			
				Classroom teaching experience	Teaching in First Grade	Teaching in the school district	Teaching in current school
Total sample	84	B = 84 M = 33 P = 21	E = 84 B = 8 S = 2 R = 1 A = 2 O = 4	10.64 (9.73)	4.86 (6.04)	8.85 (9.00)	6.62 (7.22)
TSG	40	B = 40 M = 14 P = 6	E = 40 B = 4 S = 0 R = 0 A = 1 O = 2	11.66 (9.69)	5.45 (6.12)	8.97 (8.46)	6.37 (6.88)
Comparison	44	B = 44 M = 19 P = 15	E = 44 B = 4 S = 2 R = 1 A = 1 O = 2	9.74 (9.80)	4.35 (6.00)	8.74 (9.54)	6.83 (7.58)
Year 1	34	B = 34 M = 9 P = 8	E = 34 B = 5 S = 2 R = 1 A = 2 O = 0	10.29 (9.85)	3.88 (5.01)	8.03 (8.00)	5.76 (5.53)
CA TSG	14	B = 14 M = 3 P = 2	E = 14 B = 1 S = 0 R = 0 A = 1 O = 0	10.00 (10.78)	5.14 (7.29)	6.36 (7.48)	5.14 (6.27)
CA Comparison	20	B = 20 M = 6 P = 6	E = 20 B = 4 S = 2 R = 1 A = 1 O = 0	10.50 (9.43)	3.00 (2.27)	9.20 (8.33)	6.20 (5.07)

Year 2	50	B = 50 M = 24 P = 12	E = 50 B = 3 S = 0 R = 0 A = 0 O = 4	10.89 (9.74)	5.57 (6.66)	9.45 (9.70)	7.23 (8.23)
TSG	26	B = 26 M = 11 P = 4	E = 26 B = 3 S = 0 R = 0 A = 0 O = 2	12.63 (9.09)	5.63 (5.49)	10.50 (8.77)	7.08 (7.23)
Comparison	24	B = 24 M = 13 P = 8	E = 24 B = 0 S = 0 R = 0 A = 0 O = 2	9.09 (10.27)	5.52 (7.82)	8.35 (10.65)	7.39 (9.32)
CA TSG	11	B = 11 M = 4 P = 1	E = 11 B = 3 S = 0 R = 0 A = 0 O = 1	14.95 (9.69)	6.55 (5.80)	14.18 (9.94)	7.55 (8.13)
CA Comparison	10	B = 10 M = 5 P = 5	E = 10 B = 0 S = 0 R = 0 A = 0 O = 1	9.40 (11.17)	6.10 (7.65)	9.70 (12.26)	8.10 (9.10)
PA TSG	10	B = 10 M = 4 P = 2	E = 10 B = 0 S = 0 R = 0 A = 0 O = 1	12.75 (8.50)	6.10 (5.49)	8.90 (6.71)	8.00 (7.16)
PA Comparison	10	B = 10 M = 5 P = 3	E = 10 B = 0 S = 0 R = 0 A = 0 O = 1	9.10 (11.43)	5.50 (9.42)	7.90 (11.04)	7.60 (11.17)
VA TSG	5	B = 5 M = 3 P = 1	E = 5 B = 0 S = 0 R = 0 A = 0 O = 0	3.67 (2.31)	0.67 (0.58)	2.33 (1.15)	2.33 (1.15)

VA Comparison	4	B = 4 M = 3 P = 0	E = 4 B = 0 S = 0 R = 0 A = 0 O = 0	8.00 (3.00)	3.67 (1.53)	5.33 (0.58)	4.33 (2.08)

CA= California; PA=Pennsylvania; VA=Virginia.
B=Bachelor; M=Masters; P=Post Masters.
E = elementary Ed (this includes the multiple categories from CA); B = Bilingual Education; S = Special Educatio; R = Reading Specialist ; A = Administrative; O= Other

Appendix B

Fidelity Checklist: TSG Audiotape Recordings

School: _____ Facilitator: _____ Date: _____

Lesson number (circle one): 1 2 3 4 5 6 7 8

Lesson was covered in meeting(s): _____ Lesson Topic: Vocabulary _____ Comprehension _____

Section Title	**Description**	**Check One**	
		NO *Not Done*	**Yes** *Done*
1. Debrief	Asks teachers to share their experiences applying last session's strategy, by (check all that apply): • Asking what went well, what did not _____ • Asking how students responded _____ • Other _____ _____		
2. Cover Key Points in the Research	Reviews the material assigned for the session, by (check all that apply): • Asking teachers to share the highlights _____ • Asking teachers target questions _____ • Summarizing the material _____ • Linking the material to examples _____ • Other _____ _____		
3. Analyze Upcoming Lesson According to the Research Principle	Collaboratively discusses the upcoming lesson, by (check all that apply): • Discussing how lesson does or does not reflect research principles from research material _____ • Discussing strengths & weaknesses of the lesson _____ • Other _____ _____		
4. Collaboratively Plan an Upcoming Lesson	Assists collaborative planning of a common lesson, by (check all that apply): • Having teachers work together to plan an upcoming lesson _____ • Providing expertise _____ • Facilitating collaboration _____ • Other _____ _____		

Provide Comments for Each of the Following:

Level of Participation:

_____

Level of Participant Preparation:

_____

Pace of TSG: _____

Facilitator Debrief Comments:

General Comments: