Running head: Teacher Study Group

# Teacher Study Group: Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms

Russell Gersten

Instructional Research Group

Joseph Dimino

Instructional Research Group

Madhavi Jayanthi

Instructional Research Group

James S. Kim

Harvard Graduate School of Education

Lana Edwards Santoro

Instructional Research Group

Citation: Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (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.

# Abstract

Randomized field trials were used to examine the impact of the *Teacher Study Group* (TSG), a professional development model, on first grade teachers' reading comprehension and vocabulary instruction, their knowledge of these areas, and on the comprehension and vocabulary achievement of their students. The multi-site study was conducted in three large urban school districts from three states. 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 also significantly outperformed control teachers on the teacher knowledge measure of vocabulary instruction. Confirmatory analysis of student outcomes indicated marginally significant effects in oral vocabulary.

# Teacher Study Group: Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms

Over the past thirty years, a body of research on promising practices for effective professional development (PD) has slowly emerged (e.g., Berman & McLaughlin, 1978; Huberman & Miles, 1984; Ball, 1990; Garet, Porter, Desimone, Birman, & Yoon, 2001). Although this body of research has had a profound impact on the field, a good deal of uncertainty remains. Advocated PD practices, though sensible and compelling, have rarely been widely field-tested and evaluated using rigorous research techniques (Guskey, 2003; Desimone, 2009). The majority of the studies in PD encompass a broad array of methodologies (surveys, comparative case studies, qualitative, mixed methods) that rely heavily on teacher self-report and case study analysis. We possess very little empirical evidence on the critical role these promising practices play in enhancing teacher learning of effective instructional strategies, and more importantly student learning (Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Rarely do studies link PD to student outcomes, and the few that have attempted to do so have often yielded disappointing results. In fact, Garet et al. (2008) found that large-scale PD that included many of these practices did lead to significant increases in teachers' knowledge and observed teaching practice, but failed to enhance student reading achievement significantly.

To address this need for effective PD, we developed a PD program, the *Teacher Study Group*, and tested its effects on teacher and student outcomes using randomized control trials. In this article we delineate the features and format of the PD program, and describe the multi-site randomized control trials study.

## Teacher Study Group

The *Teacher Study Group* (TSG) PD program used in this study was based on our earlier research efforts to identify strengths and weaknesses of PD strategies for translating research into teaching practice in high poverty schools (Gersten, Morvant, & Brengelman, 1995; Gersten & Brengelman, 1996; Gersten, Darch, Davis, & George, 1991; Gersten & Woodward, 1990) as well as principles gained from reviews of the PD literature. The TSG PD program was then defined and operationalized by senior research associates on the research team. As the name suggests, the TSG PD program employs teacher study groups, a form of professional development.

To the best of our knowledge, the term "teacher study group" first appeared in the literature almost 25 years ago (Sugai, 1983). Since then, the term teacher study group has referred to a rather loose conglomerate of PD approaches (Logan & Stein, 2001; Taylor & Pearson, 2003; Tichenor & Heins, 2000) that have very little in common, except for comprising of small groups of teachers working together towards a specific goal. In 1992, Murphy identified three main purposes for teacher study groups in general: (a) facilitate implementation of curricular and instructional innovations, (b) plan school improvements, and (c) guide educators in studying research-based practices. This broad, virtually all encompassing definition seems to fit the literature on teacher study groups we reviewed.

Although teacher study groups have been used sporadically for the last two decades, very few details have emerged about their specific features and their impacts on either teaching practice or student outcomes. Findings are only suggestive of the link between these groups and improvements in teaching practice, student achievement, and school culture (e.g., Foorman & Moats, 2004; Gersten, Baker, & Griffiths, 2003; Saunders et al., 2001; Tichernor & Heins, 2000). For example, Foorman and Moats used the teacher study group as one of several components of their PD work in Washington, D. C. on improving the quality of reading instruction in schools and indicated that both the research team and the teachers found them to be very promising. Saunders et al. noted that teacher study groups were an essential component of their successful school-wide improvement program, which resulted in documented achievement growth in reading. In our case study research of the Early Literacy Project (Englert & Tarrant, 1995), we found teacher study groups to be linked to high levels of change in teacher beliefs and use of curricula, though not necessarily to shifts in the use of research-based instructional principles in reading (Gersten et al.).

In a departure from the above examples, Tichenor and Heins (2000), by means of a design experiment, examined the use of teacher study groups to explore ways to meet the needs of at-risk students. In this design example, the teacher study groups were the only component of the PD offering. The teacher study groups met once a month for nine months to discuss readings on research-based strategies for teaching at-risk students and to determine ways in which the research strategies can be integrated into their classroom teaching. The authors indicated that teachers in the study group responded favorably to the experience and noted a change in their student participation and self-esteem.

Essential to note from these examples, is that empirical evidence linking the teacher study group approach to teaching practice or student learning is non-existent. By clearly articulating the components of our TSG PD program and field-testing its effectiveness, we will be adding to the research base on the teacher study group approach. Given the mandates of the No Child Left Behind Initiatives that teachers have access to high-quality, research-based PD in reading, and the paucity of rigorous research in PD in general (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), the importance of adding to the research base in PD needs to be underscored.

*Conceptual Framework of the TSG PD Program used in the Study*

The TSG PD program is grounded in the research on translating research into practice (e.g., Gersten et al., 1995; Gersten & Brengelman, 1996; Gersten & Woodward, 1990) and on the promising best practices of PD that have emerged over the years (e.g., Birman, Desimone, Porter, & Garet, 2000; Garet et al., 2001; Huberman & Miles, 1984; Wiley & Yoon, 1995). As we designed the TSG program, we attempted to orchestrate major trends in PD research – integrating conceptual understanding and practical application, sustaining active learning, nurturing collegial support networks, and maintaining coherence with existing demands – into a feasible model for use in elementary schools. See Figure 1 for a pictorial representation of the conceptual framework underlying the development of the TSG.

*Integrating conceptual understanding and practical applications.* Research suggests that PD works best when it attempts to integrate conceptual understanding with the pragmatic, procedural aspects of teaching practice (Cohen & Hill, 2000; Kennedy, 1999; Garet et al., 2001). The goal of the TSG PD program was 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 program was not to change a district's core curriculum, but to *enhance* implementation of that curriculum (Gersten & Brengelman, 1996) by using research based strategies that may not be included in the teacher's guide. Teachers in the TSG program learn research-based instructional techniques that are practical and easy to implement. By having teachers prepare lesson plans that take into account students' curricular and individual needs, the TSG program links training to job-specific challenges and is a more meaningful, useful, and motivating experience for teachers.

*Sustaining active learning.* Many think that effective PD should be extensive, intense, and sustained to support change, as teachers need active, on-going learning opportunities to integrate learning into daily practice (Gersten, Chard, & Baker, 2000; Garet et al., 2008; Goldenberg & Gallimore, 1991; Showers, Joyce, & Bennett, 1987). The TSG program was designed to provide PD over an extended period of time. Teachers participate in multiple sessions and have on-going opportunities to discuss and apply research-based concepts to their classroom instruction. Because teachers actively engage in problem-solving discussions and applied learning activities, they are better able to apply their learning to their classroom teaching (Gersten, et al.).

*Nurturing collegial support networks.* PD efforts with supportive, collaborative environments are considered to be effective in bringing about change (Buysse, Sparkman, & Wesley, 2003; McLaughlin, 1994; Talbert & McLaughlin, 1994), the assumption being that teachers feel supported as they discuss research-based innovative techniques and the realities of implementing those techniques in their classrooms. Such opportunities usually include teachers from the same school and same grade level in order to provide an appropriate venue for sharing and learning. Essential to the TSG PD program are the collaborative interactions and collective participation of its teacher participants. As teachers engage in interactive discussions, it helps build common understandings of educational research. The TSG program facilitates development of collegial support networks as it includes facilitator-guided discourse and inquiry, rather than a "top-down" or "expert" led study group where teachers play a passive role listening to instruction or watching demonstrations.

*Maintaining coherence with existing demands*. Coherence is often described as the extent to which the PD program is aligned with state and district standards and the extent to which the PD experience is part of an integrated system of teacher learning (Birman et al., 2000). In recent research on the dimensions of PD, teachers identified coherence as the dimension with the strongest effects (Desimone, Garet, Birman, Porter, & Yoon, 2003; Garet et al., 2001). TSG sessions offer coherence by helping teachers make sense of the priorities that are associated with school, district, and state initiatives, by teaching research-based strategies advocated by each group and providing opportunities for teachers to collaboratively plan lessons using those strategies.

In summary, with the inclusion of these promising features of high quality PD, our TSG approach was designed to address many of the shortcomings of the more popular and widely-used forms of PD such as workshops and summer institutes (Klinger, Vaughn, & Schumm, 1998; Moss et al., 2008).

### *PD Approaches Similar to TSG Program*

Our TSG PD program shares many common features (e.g., school-based, collaborative planning, and collective participation) with other similar PD approaches that employ small groups of teachers such as lesson study and learning communities. To accentuate the uniqueness of the TSG program, we discuss these two approaches briefly and draw attention to some of the similarities and differences between these approaches and our TSG PD program.

*Lesson study.* Lesson study *(jugyoukenkyuu)* is a translation of the Japanese words *jugyou* (instruction, lessons, or lesson) and *kenkyuu* (research study) (Lewis, Perry, & Murata, 2006). Lewis et al. note that lesson study is a "large family of instructional improvement strategies" with one shared feature – "*observation of live classroom lessons by a group of teachers who* *collect data on teaching and learning and collaboratively analyze it*" (pg. 3, emphasis in original). The most salient features of lesson study involve collaborative lesson planning and implementation of live classroom lessons.

Most examples of lesson study involve mathematics instruction, with a few extending into writing and history instruction (Blum, Yocom, Trent, & McLaughlin, 2005; Lewis, Perry, Hurd, & O'Connell, 2006). While there are no experimental studies involving lessons study, several case studies exist showing promise of the lesson study approach (e.g., Lewis, Perry, & Hurd, 2009; Lewis et al., 2006; Blum et al., 2005). Lewis and her colleagues note that lesson study improves instruction by refining lessons and strengthening teachers' knowledge, commitment, community, and resources. They also report significant gains in students' mathematics scores, without claiming a causal connection between student achievement and their school-wide lesson study PD offering (Lewis et al., 2006).

In a typical lesson study, a group of teachers discuss and then collaboratively plan a common "research lesson". After collective discussion and planning, one teacher agrees to teach the "research lesson" while the other teachers observe and take notes. A post-lesson meeting is held to reflect on the strengths, weaknesses, and challenges of the lesson's implementation. The lesson study process continues with subsequent sessions in which teachers share lesson development responsibilities and take turns teaching the "research lessons." The term "research lessons" is often used with lesson study due to the process of "trying out" and revising lessons based on teacher action research (e.g., observations and classroom notes).

Like lesson study, our TSG PD program is school-based and includes collaborative lesson planning and discussions about implemented classroom lessons. Unlike the lesson study approach where experiences with "research lessons" are used to generate discussion and planning of the next "research lesson," our TSG sessions are structured around a pre-planned sequence of topics. The TSG approach relies on empirical research to inform discussions and decisions about instruction, and uses teachers' craft knowledge as a means to translate research into refined lesson plans. Lesson study, in contrast, does not focus on empirical research findings. Another distinction between lesson study and TSG PD program is the use of teacher observations of collaboratively planned lessons. We thought such formal peer-observations might negatively impact the establishment of collegial support networks. Formal teacher observations could potentially induce too much anxiety and unnecessary pressure. As a result, our TSG PD program does not to include peer-observations.

*Learning communities*. The term *learning community* refers to many different models of collaborative groups, such as Professional Learning Communities (PLCs), teacher communities, critical friends groups, and communities of practice. In a recent review of literature on learning communities, Vescio, Ross and Adams (2008) note that "all combinations of individuals with any interest in schools are now calling themselves PLCs". Although the name and purpose of learning communities vary, these groups often share a focus on learning, collaboration, and reflective dialogue.

There is some empirical evidence suggesting that learning communities can foster teacher learning (Borko, 2004; Buysse et al., 2003; Englert & Tarrant, 1995; Little, 2002; Wilson & Berne, 1999) and improve the professional culture of a school (Vescio, Ross, & Adams, 2008). There are a few studies, none with rigorous designs, on the impact of learning communities on student learning (Guskey, 1997; Vescio et al.). This emerging research suggests that collaborative groups focusing on improvements in actual teaching practice, rather than improvements in teachers' content knowledge, may have some impact on student outcomes.

Typically, teacher learning communities consist of a team of teachers that meet regularly to learn new topics, share ideas, and problem solve. The teacher teams determine the topics they would like to learn and the way they will gain this new knowledge. They might read articles or books, ask experts to speak to the group, or attend trainings in the area. Learning communities usually share an interest in improving student achievement by improving their own teaching practice. This shared interest brings coherence and continuous learning to their professional development (Vescio et al., 2008).

The TSG PD program shares key similarities with the learning community model. They both aim to develop a collaborative schoolwork culture that encourages teachers to share expertise with each other (Lieberman & McLaughlin, 1992; Sarason, 1972, Wenger, 1998; Thompson, Gregg, & Niska, 2004). In both approaches, teachers interact with each other to expand professional knowledge and reflect on their teaching practice. Unlike learning communities, which can include teachers from across the district, the TSG program is school-based and grade-specific. Learning communities also lack the focused, research-based scope and sequence found in our TSG approach. Teacher teams in the learning communities can choose what they would like to read or attend and change their focus at any time. They may or may not choose to study research-based strategies. In the TSG program, teachers learn a set of research-based strategies in a sequential and logical manner, so that research is easily translated into practical, easy-to-implement strategies that are more likely to improve reading instruction, and thus student learning.

### *Need for Effective PD in Reading Comprehension and Vocabulary*

The improvement of comprehension and vocabulary instruction is viewed by many as an urgent priority in the field of reading (e.g., Duffy, 1993; Block, Gambrell, & Pressley, 2002).

However, the PD offerings that are often available to the teachers do not address this need. Most PD efforts in reading have been in the areas of phonemic awareness, decoding, and fluency, rather than in comprehension or vocabulary (Moss et al., 2008; Haager, Heimbichner, Dhar, Moulton, & McMillan, 2008). Recent survey data from the Reading First Implementation Evaluation Study indicate that teachers feel inadequately prepared to teach comprehension and vocabulary to their struggling learners (Moss et al., 2008). Given the shortcomings of the current PD offerings, the need for effective, high quality PD in reading comprehension and vocabulary is clear.

Most reading researchers concur that explicit, highly interactive instruction is essential for quality comprehension and vocabulary instruction (e.g., Beck, McKeown, & Kucan, 2002; Carlisle & Rice, 2002; Dole, 2003). Although we now see a good deal of material on comprehension and vocabulary instruction in core reading programs, the type of interactive, scaffolded instruction recommended by reading experts rarely is found in contemporary American classrooms (Pressley & Wharton-McDonald, 1998; Durkin, 1978, James-Burdamy et al., 2009). The TSG PD program was designed to address this need by helping teachers *enhance* implementation of their reading curriculum using research-based instructional strategies for teaching comprehension and vocabulary.

### *Purpose and Research Hypotheses*

The goal of this study was to evaluate the impacts of the TSG PD program on teacher knowledge of relevant research in comprehension and vocabulary instruction, classroom applications of these research-based strategies, and student reading outcomes. We designed the TSG as a concerted PD approach to build teachers' "working knowledge" of research-based comprehension and vocabulary instruction, to improve their teaching practice, and to increase student achievement in reading. We hypothesized that the program's inclusion of promising features of effective PD (e.g., Dole, 2003; Goldenberg & Gallimore, 1991; Elmore, 2002) would create gains at the teacher level, while the inclusion of elements of scaffolded instruction (e.g., Klinger et al., 1998) and a focus on key comprehension and vocabulary strategies (e.g., Rosenshine & Meister, 1994; Beck et al., 2002, Baumann & Kame'enui, 2004) would create gains at the student level.

We describe a hypothesized causal pathway in Figure 1. We hypothesized that, by participating in TSG sessions, teachers would not only learn research-based principles for vocabulary and comprehension instruction but also, through ongoing opportunities to review, reflect on, and discuss, apply these research concepts to the curriculum being used in their school. We hypothesized that this emphasis on both conceptual and procedural understanding of key principles from the research on comprehension and vocabulary instruction through ongoing collaborative learning will bring about changes in teacher knowledge and what we have called teachers' working knowledge of these principles (Gersten & Woodward, 1992).

Showers et al. (1987) note, "what a teacher thinks about teaching [practices] determines what the teacher does in the classroom" (p.85), which has been supported by a small body of research over the last 20 years (e.g. Hill, Rowan & Ball, 2005; Phelps & Schilling, 2004; Baker, Gersten, Dimino, & Griffiths, 2004). As teachers review research principles and gain insight into the practical applications of the research base and its potential benefits, we hypothesized that increased knowledge will be a strong motivator for changes in daily teaching practice.

The *Teacher Study Group* program is designed to support the improvement of teacher knowledge and teaching practice in ways that improve student achievement. As teachers use their "working knowledge" (Gersten & Woodward, 1992) by consciously improving their lesson plans and their instruction based on empirical research, we expect corollary gains in student performance. Thus, we hypothesized that improvements in teacher knowledge and teaching practice will facilitate improvements in student reading comprehension.

## Method

### *Setting and Participants*

The multi-site study was conducted in three large urban school districts. Sites were in three states: California, Pennsylvania, and Virginia. A total of 19 schools were involved in the study (10 TSG, 9 control). All schools were involved in the Reading First program, which entailed a good deal of support for professional development in reading for primary grade teachers.

The initial baseline teacher sample included 84 first grade teachers (40 TSG, 44 control); however, three teachers (1 TSG, 2 control) dropped out of the study. Each teacher left for a different reason: family problems, illness, and moving from the school district. The final analytic teacher sample consisted of 81 teachers (39 TSG, 42 control).

Seven students were randomly selected from each class to examine the impact of the Teacher Study Group. Our initial student sample included 575 students[i] (273 TSG, 302 control), with mobility resulting in a final analytic sample of 468 students (217 TSG, 251 control).

*Teacher demographics*. The teacher demographic data are summarized in Table 1.Of the 39 first grade teachers in the TSG group, only three were male and one-third possessed a master's degree in education. For control group teachers, only four were male, and 42.86% earned a master's degree. TSG teachers had, on average, 11.35 years of classroom teaching experience (SD = 9.63; range 1 - 31 years) and 5.05 years of experience teaching first grade (SD = 5.70; range 0 - 23 years), whereas control group teachers had 9.59 years of classroom teaching experience (SD = 9.76; range 0 - 36 years) and 4.39 years of experience teaching first grade (SD = 6.14; range 0 - 32 years). None of these differences were statistically significant. However, the difference between TSG and control groups in teachers with education beyond a master's degree was marginally significant ($\chi_2$ (1, 81) = 2.82, $p$ = .093), favoring the control group.

*Student demographics*. Student demographic data by site 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 site to site we chose to use the more inclusive term and provide descriptive data on student's scores in Table 2.

*Baseline equivalence of schools*. 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 3, 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.

*Student attrition*. The overall attrition of the student sample was 18.6%, while the differential attrition between the TSG group and the control group was 3.6%. These attrition rates are considered acceptable levels for a RCT (What Works Clearinghouse, 2008). 20.5% of students from the TSG group (n = 56) and 16.9% of students from the control group (n = 51) were unavailable for post testing, primarily due to family relocation. Chi-square analysis revealed no significant relationship between the proportion of children with missing data in the TSG group and the control group, $\chi^2$ (1,575) = 1.24, $p$ = .265.

We also 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 three Dynamic Indicators of Basic Early Literacy Skills (DIBELS) measures (Good & Kaminski, 2002) – *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 *Woodcock Diagnostic Reading Battery (WDRB)* measures of *Reading Vocabulary* (*t*(573) = 1.06, *p* = .291) and *Passage Comprehension* (*t*(573) = .57, *p* = .567). These results suggest that differential attrition did not impact the external validity of the study.

*TSG facilitators.* There were five TSG facilitators: two in the California district, two in the Pennsylvania district, and one for the Virginia district. The facilitators had a strong background in reading research and means for translating this body of research into classroom practice. Four had doctoral level degrees in special education or literacy and experience with reading research. The fifth facilitator had extensive district administration experience and a background in reading instruction.

*Design*

A randomized controlled field trial was used to examine the impact of the TSG intervention. In Year 1 (2004-2005) the study was conducted in a school district in California. In year 2  (2005-2006) the study was replicated in school districts in Pennsylvania and Virginia; additional schools in the California site participated. 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 California school district and Pennsylvania school district, schools were matched prior to random assignment. In the California district, 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 Pennsylvania 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). The sample in the Virginia district included three schools. All teachers and schools in the study were remunerated for their participation.

### *TSG and Control Conditions*

The Reading First grants administered in the three districts mandated that all teachers in Reading First schools allocate certain time for PD efforts that were focused on scientifically based reading approaches. However, districts had wide latitude in how they operationalized and implemented these PD activities in reading. Teachers from all three districts attended a summer institute in reading and met during the year for the contracted PD efforts on reading, that were mandated under Reading First. In the school districts in Pennsylvania and California, participation in TSG counted towards the required PD hours. In Virginia, it was as add-on.

PD activities attended by teachers from the TSG group and the control group during the school year are summarized in Table 4. 77% of the teachers from TSG and 57% of the teachers from control attended PD activities in comprehension. PD activities in vocabulary were attended by 29% of the teachers from TSG and 23% from control. Chi-square tests revealed no significant differences between the groups for the various PD activities attended by teachers, except for PD on Structured English Immersion Techniques ($\chi 2$ (1, 81) = 11.82, $p$ = .001), which was attended by more teachers from the TSG group.

The reading curriculum used in TSG and control classrooms was a constant within each school district. *Open Court Reading* (Adams et al., 2000; Bereiter et al., 2002) was used in the California school district, *Harcourt Trophies* (Harcourt School Publishers, 2005) in Pennsylvania school district. In the Virginia school district, teachers used a guided reading approach with no core text; however, the Wright Group's *Sunshine Reading Series* (1996) was a source as was *Units of Study for Primary Writing* (Calkins & Mermelstein, 2003), along with trade books. Guided reading entailed small group reading instruction with students grouped by reading ability using leveled reading materials. Guided reading lessons were typically 15 to 20 minutes long. Teachers also worked on word study and writing, often using whole class instruction.

*Control condition.* Teachers in the control condition participated in scheduled school and district PD activities. During the study, control teachers did not participate in the TSG sessions or have access to the materials. After the Year 2 study ended, TSG facilitators helped implement TSG sessions in control schools in Pennsylvania and Virginia who had expressed interest in implementing the TSG in their school districts. Although the option was offered in California, no schools accepted this option, in part because of the large number of other PD offerings provided by the state and district at that time as part of Reading First.

*TSG intervention.* The TSG intervention was comprised of 16 interactive sessions held at the school site twice a month from October to mid-June; each session was approximately 75-minutes in duration[ii]. 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 PD 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.

Of the 16 sessions, the first eight focused on vocabulary instruction. The scope and sequence of the vocabulary sessions was based on *Bringing Words to Life: Robust Vocabulary Instruction* (Beck et al., 2002). These sessions addressed developing student friendly definitions, selecting words to teach, techniques for introducing vocabulary, incorporating activities to ensure multiple meaningful exposures to new words, determining the meaning of words in context, and creating a rich vocabulary environment. The eight comprehension sessions focused on the rudiments of explicit reading comprehension instruction, question-answer-relationship strategy, generating main ideas, making and evaluating predictions, and story grammar instruction (e.g., Duffy, 2002; Raphael, 1986; Vaughn, S. & Linan-Thompson, S., 2004).

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. A 4-step recursive process (described below) was instituted during each TSG session to provide 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.

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. 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 assigned readings. If the teachers did not readily discuss the selection or did not address the most important and relevant aspects of the material, the facilitators prompted them with specific questions geared towards discussing these issues. During segment three, *Walk Through the Lesson*, the teachers reviewed a lesson from the core reading program's Teacher's Guide that they would be teaching before the next TSG 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. In segment four, *Collaborative Planning*, teachers worked as a whole group or in pairs to plan a lesson that incorporated the targeted research principle. 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. Children's literature trade books were also provided to teachers for a vocabulary lesson planning activity.

### *Implementation Fidelity*

Fidelity of implementation was calculated using procedures outlined by Fuchs and their colleagues in their field implementation studies (e.g., Fuchs, Compton, Fuchs, Paulsen, Bryant, & Hamlett, 2005; Fuchs, Fuchs, Finelli, Courey, & Hamlett, 2004). All TSG sessions were tape-recorded. The facilitators did not know which audiotapes would be checked for fidelity. To determine implementation fidelity, our research staff listened to audiotapes from each site for Sessions 2 and 3 on vocabulary and Sessions 12 and 15 on comprehension. The research team selected these sessions as the lesson plans indicated application of all key components of the TSG program. Of the 36 tapes chosen for the fidelity check, three were unavailable for review due to missing data or audio taping malfunction.

Fidelity was assessed using a checklist. Each item on the checklist was marked as observed or not observed. Fidelity was calculated as percentage of items implemented (number of items observed divided by total number of items [observed and not observed] times 100). Fidelity means for each TSG session ranged from 83.3% to 93.8% with a mean of 86.5% of the key components fully implemented.

### *Measures*

*Teacher measures.* Teacher measures for this study include measures of teacher practice, teacher knowledge, teacher perception of professional culture, and teacher appraisal of the TSG intervention. Measures of teacher practice and teacher knowledge were the outcome measures for confirmatory analyses while the remaining two were examined in exploratory analyses.

We used the *Reading Comprehension and Vocabulary (RCV) Observation Measure* (Gersten, Dimino, & Jayanthi, 2007) as a posttest of teaching practice in comprehension and vocabulary. We developed the *RCV Observation Measure* to assess the quality of classroom reading comprehension and vocabulary instruction. The *RCV Observation Measure* is a moderate-inference frequency measure. It helps capture a variety of effective teaching/instructional behaviors. Based on our pilot test, we found that some of these practices would be implemented by most teachers (e.g., previewing a reading selection; asking literal questions); others would be implemented by only some of the teachers and were likely to be indicators of interactive comprehension strategy and vocabulary instruction (e.g., modeling how to make inferences, and thinking-aloud a character analysis to make overt the cognitive thinking processes).

The measure is well aligned with findings from the extant literature on effective vocabulary and reading instruction for the elementary grades (e.g., Anderson, Evertson, & Brophy, 1979; Baumann & Kameenui, 1991; Beck et al., 2002; Graves, 2006). The items in the comprehension and vocabulary domains reflect two major pedagogical aspects of effective instruction: explicitness of instruction and interactivity of instruction (i.e., the amount of scaffolding practice and feedback provided) (Beck, McKeown, Sandora, Kucan, & Worthy, 1996; Pressley, 1998; Dole, Duffy, Roehler, & Pearson, 1991). Items in the *RCV Observation Measure* also tap into other aspects of classroom teaching such as overall quality of comprehension and vocabulary instruction, student engagement, classroom management and teacher responsiveness to students, and areas of problematic instruction. The measure includes frequency items, wherein the observer tallies the number of times the item (i.e., the teaching behavior) is seen, items with a Yes/No answer format, and Likert scale items. See Sample items in Appendix A.

Data is recorded in 15-minute intervals. A 90-minute classroom observation therefore translates to 6 intervals. During each 15-minute interval the observer tallies the number of times a behavior is seen and also responds to some questions with Yes/No answer format. After each 15-minute interval, the observer turns the page and continues collecting data on a "new" protocol. Observers are encouraged but not mandated to take a few field notes to help them confirm their tallying and provide examples of the instructional practices they observe. At the end of the observation, the observer completes additional items (Likert scale items and questions with Yes/No answer format) that elicit information on the overall quality of comprehension and vocabulary instruction, student engagement, classroom management and teacher responsiveness to students, and areas of problematic instruction.

We developed two scales from the pool of *RCV* items. The comprehension scale includes 34-items and has an internal consistency coefficient of .69. The 12-item vocabulary scale has an internal consistency coefficient of .70. The internal consistency coefficients of both scales are considered adequate and suitable for a dependent measure (Nunnally, 1978, p.245; Ponterotto & Ruckdeschel, 2007). A total score is calculated for each scale by summing all the tallies and scores awarded to the small number of relevant questions (Yes =1, No = 0; Likert scale items =1-3 or 1-4).

Relevant items from *Content Knowledge for Teaching Reading* assessment (Phelps & Schilling, 2004) served as a posttest to measure teacher knowledge in vocabulary (10 items) and comprehension (25 items). The *Content Knowledge for Teaching Reading* assessment has alpha reliability coefficients in the range of .68 to .81. 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. Items in both scales required teachers to respond on a Likert 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 (e.g., sample items: "Overall, my professional development experiences have included enough time to think carefully about, to try, and to evaluate new ideas." "Most of what I learn in professional development addresses the needs of the students in my classroom."). The *Quality Professional Development* scale has 9 items and an internal consistency coefficient of .93[iii]. 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[iv]. We modified the wording in these 6 items to reflect the grade level interactions that were central to the TSG intervention. For example, the item "Teachers in this school trust each other" was changed to "Teachers in this grade level trust each other".

The *Professional Appraisal of TSG Survey* was developed by the research team to gauge participant perceptions and opinions regarding the TSG experience (e.g., sample items: "What was the most difficult aspect of attending the TSG sessions?" "How does the TSG program compare with other professional development activities you have attended?"). The survey consisted of 10 items; Likert scale, open-ended, and closed-ended questions were included.

*Student measures.* Student measures included measures of early literacy skills for use as covariates and for exploratory analyses, and measures of vocabulary and comprehension outcomes for confirmatory analyses. To assess early literacy skills, students were administered three Dynamic Indicators of Basic Early Literacy Skills (DIBELS) measures (Good & Kaminski, 2002; Kaminski & Good, 1996) – *Letter Naming Fluency*, *Phonemic Segmentation Fluency,* and *Oral Reading Fluency,* and two sub-tests of the *Woodcock Diagnostic Reading Battery (WDRB) – Word Attack*, *Letter-Word Identification*. Three other sub-tests of the *WDRB – Oral Vocabulary*, *Reading Vocabulary*, and *Passage Comprehension* were the main outcome measures.

*Letter Naming Fluency (LNF)* is a 1-minute timed measure that assesses the accuracy and speed with which children identify letter names. *LNF 6th Edition* has test-retest reliability of .88, and a predictive validity of .65 for reading performance a year later. *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. *PSF* has a test-retest reliability of .88 and a predictive validity of .68 for end of first grade reading on the *Woodcock Johnson*. *Oral Reading Fluency (ORF)* is a 1-minute timed measure that assesses a child's ability to read grade level passages fluently and accurately. *ORF* has a test-retest reliability in the .90s (Good & Kaminski, 2002).

*Oral vocabulary* subtest of *WDRB* measures a student's knowledge of word meanings. Students respond with one word that is either an antonym or synonym of the word presented orally by the examiner. The median internal consistency reliability for this subtest is .90 (Woodcock, 1997). In contrast, *Reading Vocabulary* subtest measures a student's ability to supply an appropriate synonym or antonym for a word the student reads. *Reading Vocabulary* has a median internal consistency reliability of .92. *The Passage Comprehension* subtest utilizes a modified cloze procedure. Students start by pointing to the picture represented by a phrase. As the items progress, students read a short passage and identify the missing key word. The median internal consistency reliability for this subtest is .90.

The *Letter-Word Identification* subtest measures a student's symbolic learning and reading identification skills with isolated letters and words. As items become more difficult, less frequently used and irregularly spelled words are presented. The *Letter-Word Identification* subtest has a median internal consistency reliability of .94. The *Word Attack* subtest assesses a student's skill in applying phonemic and structural analysis skills to the pronunciation of pseudowords. The median internal consistency reliability for this subtest is .92.

### *Classroom Observations*

Classroom observations were conducted in each classroom during April and early May. All teachers were observed during the reading and language arts block, which typically ranged from 2.0 to 2.5 hours. All observed teachers were told that they would be observed teaching reading during their regularly scheduled reading and language arts block; they were not given any other information regarding the purpose of the observation, nor were they told that the observations were limited to comprehension and vocabulary.

All classroom observers had classroom teaching or school experience and participated in a two-day training session to learn the *RCV Observation Measure*. 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. Quality control checks for the observations were conducted during the first ten days of observations by senior members of the research team to ensure desired level of accuracy and to correct for possible errors in coding. Feedback was provided to the observer immediately following the observation. Our trained observers were blind to teacher assignment to treatment condition (TSG or control). While our observers knew that the observations were being conducted as part of a research study, they were not told about the purpose of the study.

All teachers were observed for one full reading and language arts lesson. We observed 22% of the classrooms (n=18) for two full reading/language arts lessons to estimate temporal stability. While more frequent classroom observations would provide a much more precise estimate of the nature of vocabulary and comprehension instruction in each classroom, our major goal was to estimate impacts of the intervention on mean performance of each sample. Based on our previous experiences in collecting classroom observation data (e.g., Gersten, 1999; Gersten, Baker, Haager, & Graves, 2005; James-Burdumy et al., 2009) we felt that one observation would be sufficient – although not ideal – for obtaining such an estimate.

Our analysis of data collected from teachers who were observed twice supports the use of one observation per teacher. Data show that the frequency and patterns of teaching practices observed using the *RCV Observation Measure* are reasonably consistent from day to day; that is, the data have temporal stability. We found that data from Day 1 were positively correlated with data from Day 2 (comprehension: $r = .54, p < .05$; vocabulary: $r = .68, p < .01$). Also, paired t-tests indicated that the mean number of teaching practices observed did not differ significantly from Day 1 to Day 2 (comprehension: $t(17) = .50, p = .626, d =.11$; vocabulary: $t(17) = -.08, p = .940$; $d = .01$). In general, the trends and patterns of teaching practice did not vary from Day 1 to Day 2. For example, a teacher who did more literal and inferential questions and less strategy instruction on Day 1, tended to follow the same pattern during Day 2. Likewise, a teacher who reiterated and reinforced concepts frequently and asked more literal than inferential questions on Day 1 did the same on Day 2. However, teachers might emphasize one comprehension strategy on Day 1 (e.g. generating questions) and a different one on Day 2 (e.g. summarizing or sequencing). It appeared that the nature of the day’s lesson content in the core curricula determined the specific comprehension strategies emphasized, but not the frequency of interactive comprehension instruction. For vocabulary instruction, the nature of the lessons were much more similar from day to day, and the correlation between the two days is higher than for comprehension, as anticipated.

Note that the primary purpose of the observational measure was to serve as an outcome measure, and thus the high temporal stability of the *mean* scores is critical for determining reliability for purposes of this study.

*Inter-observer Reliability*

Thirty-one teachers (38% of the sample) were observed simultaneously by two observers. Data from these observations were used to determine inter-observer reliability. Given the lack of consensus in the field on the relative merits of various approaches for calculating interobserver reliability estimates (e.g., Hayes & Hatch, 1999; Lei, Smith, & Suen, 2007), we chose to use the percent agreement method for its strong face validity and ease of interpretation (Stemler & Tsai, 2008). As some of the observed teaching behaviors had a low base rate, (more so in case of the comprehension than vocabulary) the likelihood of over-inflating agreements based on unobserved behaviors could be high. To prevent this over-inflation and present an objective picture of reliability grounded in observed classroom teaching events, we limited our calculation of inter-observer reliability to only the active 15-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 reliability calculations.

Interobserver reliability was calculated in the following manner: First, agreements and disagreements were noted for each item in the active intervals. For example, if one observer had 5 tallies for an item and the second observer had 4 tallies for the same item, it was counted as 4 agreements and 1 disagreement. Then, agreements and disagreements from all the active intervals were totaled and reliability was calculated using the following formula: agreements divided by agreements plus disagreements times 100. Inter-observer reliability was on average 84.49% for the vocabulary scale and 90.89% for the comprehension scale.

*Student 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. Our data collectors had classroom teaching or school experience. Data collectors were trained in administering and scoring student assessments in a 5-hour training session. 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*

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 involved random assignment of schools to TSG and control conditions, we employed a two-level model to estimate treatment effects on teacher and student outcomes. 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.* The fully unconditional model at Level 1 for teacher *i* in school *j* can be written as $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$ , (1), and the Level 2 model for the intercept is $ß_{0j} = \gamma_{00} + \gamma_{01}$ $(TSG)_{j} + \mu_{0j}$, (2), where $ß_{oj}$, the mean score on the teacher outcome for school *j*, is regressed on the dummy variable, 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, $\gamma_{01}$. The Level 1 and Level 2 model can be combined to form the following mixed-effects model, $Y_{ij} = \gamma_{00} + \gamma_{01}(TSG)_{j}$

+ ($\mu_{0j}$ + $\varepsilon_{ij}$), (3), where $\mu_{0j}$ is a random effect for school *j,* and $\varepsilon_{ij}$ 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 estimates on student outcomes.* The Level 1 model for student *i* in school *j* can be written as $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$ , (1), where $Y_{ij}$ is the posttest reading score for student *i* in school *j*, $\beta_{0j}$ is the mean posttest score for school *j*, and $\varepsilon_{ij}$ is the error term for student *i* in school *j*. The fully specified Level 2 model is written as $ß_{0j} = \gamma_{00} + \gamma_{01}(LNF)_j + \gamma_{02}(TSG)_j + \mu_{0j}$, (2), where $ß_{0j}$ is the posttest reading score for school *j* and predicted by a pretest covariate, the school mean scores on *Letter Naming Fluency (LNF)*[v] measure, 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 $\gamma_{02}$. The Level 1 and Level 2 equations can be combined to form a mixed-effects model, which can be written as $Y_{ij} = \gamma_{00} + \gamma_{01}(LNF)_j + \gamma_{02}(TSG)_j + (\mu_{0j} + \varepsilon_{ij})$, (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. The third equation was used to estimate treatment effects on student outcomes.

*Estimated Power for Detecting Effects*

Using Optimal Design (Raudenbush, Liu, Spybrook, Martinez, & Congdon, 2006), prior to the conduct of the study, we estimated the minimum detectable effect size (MDES; Bloom, 2005), which is defined as the smallest true impact that can be detected with 80% power using a two-tailed test with alpha set at .05. (See Table 5.) We estimated the MDES to be .35 for student outcomes and .65 for teacher outcomes. For the a priori power analysis, we used the following design parameters: number of schools ($J$ = 19), estimated number of students per school ($n$ = 20), estimated number of teachers per school ($n$ = 5), anticipated intra-class correlation ($\rho$ = .05), and the percentage of the variance in the posttest explained by the covariate ($R^2$ = .65, for student outcomes only as there was no pretest covariate at the teacher level). Thus, the study was designed to detect impacts that are somewhat larger than what are typically seen in large-scale randomized control trials and evaluations (e.g., James-Burdemy et al., 2009). From the onset, we were aware that the power at the student level was low, however the funding agency limited the budget so that this was the maximal sample we could use for the research.

## Results

### *Confirmatory Analyses*

#### *Teacher Outcomes*

Results from the multilevel models used to estimate the TSG treatment effects on teacher outcomes of observed teaching practice, teacher knowledge of comprehension and vocabulary instruction, and teacher perceptions of professional culture are presented in Table 6.

*Impact on observed teaching practice.* The *RCV Observation Measure* was 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.[vi]

*Impact on teacher knowledge of reading instruction.* Impact data on the *Content Knowledge for Teaching Reading* used to measure teacher knowledge in comprehension and vocabulary instruction indicate that teachers in the TSG schools scored higher on the measure of comprehension knowledge by .32 standard deviations; however, this standardized mean difference was not significantly different from zero. The effect on knowledge of vocabulary 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.[vii]

*Impact on teacher perceptions of professional culture*. The *Quality of Professional Development* scale and *Teacher-Teacher Trust* scale were used to measure teacher perceptions. (Note that the items Teacher-Teacher Trust scale were adjusted to address trust among grade level peers rather than the school at large since the TSG is implemented at the grade level.) Each of the multilevel models also includes a pretest score on the measure, which served as a covariate and improved the precision of the estimated treatment effect. Our findings suggest that teachers in the experimental condition performed at a level that was marginally significantly different than the control group on measures of the quality of professional development ($d$ = .39). 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 posttest measures of reading achievement from the *WDRB* and *DIBELS*. See Table 7. The battery of tests included three measures that were directly related to the focus of the intervention: *WDRB Reading Vocabulary, WDRB Oral Vocabulary*, and *WDRB Passage Comprehension* and another three that addressed other aspects of reading: *WDRB Letter Word Identification, WDRB Word Attack,* and *DIBELS*

*Oral Reading Fluency.* 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. There was no statistically significant impact on measures of *Letter-Word Identification, Word Attack,* and *Oral Reading Fluency*, that is, the non-target outcomes. The magnitude of these treatment effects ranged from .13 to .23. Note that none of these aspects of reading were foci of the TSG intervention.

### *Post Hoc Power Estimates*

We conducted a post hoc power analysis using the actual parameters from the study and the intraclass correlations *($\rho$)* and the correlations between the pretest covariate and post tests ($R^2$) calculated from the data obtained from the study. (See Table 5.) We estimated a minimum detectable effect size (MDES) of 0.76 for teacher outcomes, based on 19 schools, 81 teachers, $\alpha$ =.05, power = .80, $\rho$ = .08, and *no covariate.* Likewise we estimated a MDES of 0.57 for student outcomes when based on 19 schools and the 468 students in the study, $\alpha$ =.05, power = .80*,* $\rho$ = .22, and $R^2$ = .39. The actual power for detecting our effect sizes of 0.86 for comprehension and 0.58 for vocabulary was 94% and 67% respectively. However for student outcomes in vocabulary and comprehension, the targets of the research study with effect sizes ranging from .13 to .44, we were powered, on average, at 28%. Note that the $R^2$ for student outcomes was lower and the ICC was higher than we had originally estimated. Thus, power was less than estimated in advance.

## *Exploratory Analyses*

### *Examining Correlations Between Teacher and Student outcomes*

We also conducted correlational analyses to explore the relationship between the teacher observation and knowledge measures and the average reading performance level of children in their classrooms. Since classes began with varying level of initial proficiency in reading and reading-related skills, we partialled out initial pretest scores on *Letter Naming Fluency (LNF.)* Table 8 displays these partial correlations.

We found several significant, moderately sized partial correlations between the teacher measures and student reading outcomes. The teacher knowledge measures of comprehension and vocabulary were significantly associated with all *WDRB* measures and *DIBELS Oral Reading Fluency*. Scores on the teacher observation scale for both comprehension and vocabulary were significantly correlated 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.

### *Examining Site-Specific Effect Sizes for the RCV Teacher Observation Measure*

We also explored whether effect sizes on the measure of observed teaching practice were larger for sites with higher fidelity scores. Although the fidelity of implementation score for implementation of the TSG sessions was higher, on average, in the southern California site (95%) than the Pennsylvania (75%) and Virginia sites (69%), the pattern of associations was inconsistent. Table 9 displays means and standard deviations on the observational measure of comprehension and vocabulary for each of the three sites. On the comprehension measure, the effect sizes were largest in the southern California site ($d$ = 1.33) followed by Pennsylvania ($d$ = 1.12) and Virginia ($d$ = .62) sites, roughly reflecting the relative implementation levels for the TSG program in these sites. On the vocabulary measure, the effect size was largest in the Virginia site ($d$ = 1.13) followed by the southern California site ($d$ = .80) and then the Pennsylvania site ($d$ = .34). The pattern for vocabulary observation measures does not follow the levels of TSG implementation seen in the three sites. There was a good deal of variability between schools in a given site, and these analyses are simply exploratory.

*Professional Appraisal of TSG*

Overall, participants felt positive about their participation in the TSG. Most (97%) felt that the TSG was much more useful and beneficial than other forms of professional activities they experienced. 72% of the teachers felt that the TSG helped them in teaching reading. Majority of the participants noted that they would volunteer for a TSG, if one were to be held at their school on another topic (definitely volunteer – 64%; probably volunteer – 20%).

## Discussion

Much has been written about the importance of creating more active and dynamic learning experiences for teachers in which teachers are treated as professionals (Desimone, 2009; Kennedy, 1998). Teachers often work in isolation within the four walls of their classroom. This isolation gives them few opportunities to interact with other teachers and even fewer opportunities to work together as a group to address and solve problems (Fullan, 2001; Griffin, 1995).

Although researchers and scholars have been addressing this issue for 30 years (e.g., Warren-Little, 1982; Lortie, 1975), we are only beginning to understand methods for collaborative work that actually are effective, that is, those that can significantly alter teachers' knowledge and actual teaching practice, and also show some impact on relevant student outcome measures. We developed the collaborative TSG PD program to address this issue, borrowing not only from contemporary thinking on the importance of collaborative PD efforts, but also on our understanding of the literature on how to best help teachers successfully translate research into day-to-day practice. We examined the impact of the TSG program using randomized control trials. In the ensuing paragraphs, we discuss the findings from the study and highlight the lessons we learned while conducting the multi-site randomized trials.

### *Reflections on the Findings from the Study*

*Impact on observed teaching practice and teachers' knowledge and understanding of effective practices in comprehension and vocabulary*. The study resulted in significant impacts on classroom observation scales, with effect sizes of .86 for comprehension and .58 for the vocabulary. Both scales demonstrated reasonable internal consistency (Cronbach's alpha) reliability for an outcome measure (Nunnally, 1978; Ponterotto & Ruckdeschel, 2007), suggesting that they measure coherent constructs. Inter-observer reliability was on average 84.49% for the vocabulary scale and 90.89% for the comprehension scale. Data indicate that the TSG PD program led to teachers implementing at least some of the types of interactive explicit instruction that were promoted and discussed in the sessions more frequently than for teachers not involved in the intervention.

The effect size for the comprehension scale of the observation measure was somewhat higher than that for vocabulary. The reason for this small discrepancy is unclear and it could be related to the nature of the observational measure. The key finding seems to be that teachers did significantly alter their practice in teaching of both comprehension and vocabulary.

Gains in teacher knowledge were also observed in both comprehension and vocabulary. While only the impact on vocabulary knowledge was significant ($d = .73$), impact on comprehension was in the expected direction, .32, but quite a bit smaller. 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. This iterative procedure appeared to foster greater understanding of the vocabulary material covered in the TSG program than for the material in comprehension (Beck et al.).

As we could locate no comparable book in the area of comprehension, we had 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. (2002) 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; Pressley, 2002; Stanovich, 2000), but all of them seemed better suited for a graduate course than for an ongoing PD program. We do see a need for such a book to accompany PD in the area of comprehension that can be used for work with practicing teachers.

Due to the scope and sequence of comprehension instruction in core reading programs, each comprehension strategy was covered in only one TSG session. Consequently, cumulative review of the comprehension concepts was not possible. Only one activity was common to all the TSG sessions - asking participants to analyze comprehension instruction in their core reading program by responding to a consistent set of guiding questions. (e. g., Does the lesson explicitly explain what the strategy is, when it would be used, and the steps for doing the strategy? Does the lesson provide scaffolded practice, with students having multiple opportunities to practice, gradually moving to independent strategy use?).

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 as much as their vocabulary information. 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 build a coherent understanding. This approach is consonant with that espoused by DeWitz, Jones, and Leahy (2009) who found that core reading series do approach comprehension instruction in a fragmented fashion and cover too much material superficially.

*Student outcomes*. Our study was essentially underpowered to detect significant impacts on students since these analyses are most appropriately conducted at the school level. We were aware of this shortcoming from the onset, but funding considerations limited the scope of our work. Furthermore, the post hoc power analyses revealed that our original power estimates were imprecise due to our lower estimate of the intra-class correlation ($\rho$) for the student outcomes, and a higher estimate of the correlation between the prestest covariate and post test ($R^2$) (see table 5).

However, we did find marginally significant effects for *Oral Vocabulary* ($d$ = .44, $p$<.10) on the Woodcock Diagnostic Battery. This effect size for *Oral Vocabulary* 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 changes in teaching practice in the areas of vocabulary and comprehension, but their limited reading proficiency may have dampened performance on reading vocabulary items. The impact on *Passage Comprehension*, a cloze measure, was non significant ($d$ = .13). In contrast, the effect size on the California Achievement Test in reading comprehension was .20, over fifty percent higher, although this measure was only administered in the Pennsylvania and California sites, which comprised 89% of teachers and 88% of students.

Effects were .21 and .23 for the *Letter-Word Identification* and *Oral Reading Fluency* subtests. Neither is significant but both fall in line with the impact estimates from recent cluster-randomized trials of school-level interventions like Success for All (e.g. Borman et al., 2005). 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, the effect on the *Word Attack* measure, a purer phonics measure involving decoding of pseudowords, was only .13. Although this hypothesis is highly speculative, the increased work on the meanings of words, may have enhanced scores on *Letter-Word Identification*, but not *Word Attack*, since the latter measure entails only abstract pseudowords.

While we saw impacts at the teacher level, gains were much more modest at the student level. This may be due to several factors. The first is that the study was underpowered at the student outcome level, in part because of overly optimistic estimates in our power analyses, but also because we hypothesized more modest gains at the student level than the teacher level, since student effects were distal. The intervention focused on changing teaching practice and teachers' knowledge and the impacts on student reading outcomes were indirect. It is also conceivable that increases in teaching practices occurred in response to being observed. While it is difficult to "fake" effective teaching practices during all or a significant portion of the 2-2½ hour observation period, it is possible that the TSG teachers knew what was expected of them, and on the days of observation they exhibited those targeted practices.

Interviews with the TSG teachers helped us speculate further on the possible reasons for lack of significant impacts at the student level for comprehension *on the Woodcock Diagnostic*

*Reading Battery*. TSG teachers overwhelmingly noted that they found the comprehension sessions not as fruitful for classroom use, because (a) too many comprehension strategies were covered in a short period of time, (b) information from session to session was not cumulatively reviewed or built upon, and (c) time spent was not enough to process and comprehend how to teach their students effectively. In essence, we learned that teachers need time to build proficiencies in comprehension strategies and to apply them routinely in their classrooms as the PD is occurring, and that teachers need to see the link between teaching and student learning.

*Exploring relationships between observed teaching practices, teacher knowledge and covariate-adjusted student reading outcomes*. We also explored the relationship between teacher knowledge or instruction and student reading outcomes (using covariate-adjusted posttest reading scores). The correlations between teacher knowledge and student reading outcomes were in the range of .22 to .49 with a median value of .31; correlations between teaching practice and student reading outcomes were slightly lower, .06 to .36, median of .28.  Overall, all but one of the correlations between teacher and student outcomes were significant or approached significance. To provide a perspective for interpreting these correlations, we examined correlational data from recent, large-scale field studies such as the Evaluation of the Supplemental Reading Comprehension Interventions (James-Burdumy et al., 2009) and the Reading First Impact Study (Gamse, Jacob, Horst, Boula, & Unlu., 2008). In both these studies, reading instruction was measured using observational measures that were similar to the *RCV Observation Measure* used in this study. The Reading Comprehension study used a very similar measure, one adapted (and simplified) from the RCV measure and was conducted in 5th grade only. The Reading First Impact Study included observations in grades 1 to 3 and included a good deal of information on time spent in various activities, but did devote some attention to the nature of instructional interactions. Thus, one is highly similar to the *RCV Observation Measure,* but used with much older students; the other somewhat similar, but used with students at identical and similar grade levels.

Both sets of researchers found significant correlations between teaching practice and student reading outcomes that were of lower magnitude than those found in the current study: .07-.08 for Strategy Instruction and comprehension measures in the 5th grade study and .07 between explicit teacher instruction and SAT 10 comprehension score in the Reading First Impact Study.

Contemporary theories of teacher change (e.g., Dole, 2003; Kennedy, 1998; Yoon et al., 2007) suggest that PD interventions like the TSG should first have a positive impact on proximal outcomes such as teaching practice and teacher knowledge if student achievement is to improve in the long run. The small to modest correlations from our exploratory analysis suggest that quality reading comprehension and vocabulary instruction might lead to increased growth in student achievement. Ultimately, improvements in student reading comprehension and vocabulary knowledge may depend on designing interventions that are similar to the TSG but are more intensive and provide a good deal of support in helping teachers learn how to think aloud about the comprehension process, how to provide feedback to students that encourages them to think through the material they read, and how to use student responses to promote increasingly sophisticated dialogue amongst each other about the meaning of the text and the meanings of words they are learning. The TSG was an attempt to also link the research principles to issues that arise in the curriculum used in each school. It did appear to be an effective approach, although an approach that could be refined and enhanced even more.

### *Lessons Learned From Conducting a Multi-site RCT in PD*

We implemented the TSG PD approach in three very different school districts (two of the school districts located over 2000 miles from the "home" district) and used two facilitators who were part of the development team, one member with partial involvement in development, and two with no involvement in developing the PD materials. We have identified several high points of the project as well as implementation issues we had difficulty with to inform future research in this area.

*Broad scope.* Our testing of the TSG PD program goes beyond a basic efficacy trial, as the study exemplifies characteristics of both typical efficacy and effectiveness trials. Wayne et al. (2008) define an efficacy trial as one in which the PD is delivered by its developers in fewer settings with a small sample ranging from 5 to 44 teachers. On the other hand, they note that effectiveness trials are implemented in a variety of settings with multiple trainers. The breadth of our sample, with 19 schools and 81 teachers, reaches beyond those seen in many of the earlier studies of PD. While the findings have broader applicability than typical efficacy trials, additional research is needed in this area to inform decision making in schools.

*Bringing consistency to a fragmented system.* The TSG program attempts to address the wide array of fragmented PD experiences that were part of Reading First by addressing an issue raised by Kennedy (1991) in her synthesis of the PD research, about how PD can adequately address teachers' craving for concreteness and linkage of content to their curricula with the more abstract concepts that evolve from research. Because the teachers in our study attended training institutes or workshops in the five areas of reading emphasized in Reading First, the goal of our PD activities was to help teachers learn how to apply the research they had learned to their classrooms using their own curriculum.

*Professional approach to PD.* In our previous work, we noticed that teachers often feel anxious when coached by PD trainers and describe being observed by a coach as useless and uncomfortable (Gersten & Woodward, 1992). We attempted to address these issues by developing a professional and respectful approach to PD. The Teacher Study Group consists of all the teachers in a given grade level, who meet twice a month with a knowledgeable facilitator to discuss brief readings on relevant research and then apply this research to refine the lessons they are planning to teach the following week. Instead of being watched and rated by a literacy coach, teachers report the success or failure of their refined lesson back to the group and the group works together to address the application of the research-based strategy. This structure allows teachers to share their teaching experience with their colleagues and professionally evaluate their application of the research they have learned.

*Use of formative assessment data.* The need for PD efforts to focus on improving student learning and understanding, and not just on teacher instruction is a point emphasized in some of the PD literature (American Educational Research Association, 2005; Dole, 2003; DuFour, 2004; Kennedy, 1999). Supporting this is Black and Williams' (1998) meta-analysis indicating that formative assessment data are effective in producing significant learning gains in students. While an interesting issue, the use of formative assessment data was not pursued in the current study. Our future work will begin to explore how TSG principles and practices should be shaped by formative assessment data.

*Varying levels of teacher engagement.* Discussions with the facilitators revealed varying levels of teacher engagement at all sites. Some engagement issues were related to personal problems such as child-care constraints and health concerns. Engagement did not seem to relate to years of teaching experience. Some experienced teachers took on a leadership role during the

TSG sessions, while others clearly espoused a “I know it all and there is nothing new to learn” attitude. Dealing with teachers with different levels of involvement and engagement is likely to be a reality of PD research.

*Difficulty in equating time spent in PD.* Wayne et al. (2008) noted that research on PD often compares disparate groups. Teachers in the experimental and control group rarely receive the same amount of time in PD and it is, therefore, difficult to isolate the impact of the PD approach being tested. We attempted to control for the time spent in PD across experimental and control teachers, by requesting that participation in TSG count towards the required PD hours in the school district. However, this was a difficult task to achieve. While the TSG activities counted towards mandated PD hours in school districts in California and Pennsylvania, the TSG was an add-on in the school district from Virginia. We see this issue of maintaining equal time in PD as on-going struggle while conducting large-scale impact evaluations in PD.

*Difficulty in maintaining fidelity.* Our review of implementation reinforced our sense that fidelity was more difficult to maintain at some sites 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 school districts (Pennsylvania and Virginia), TSG sessions were only implemented during 30-minute planning times. Under these circumstances, TSG sessions had to be continued across multiple planning days. Limited sessions also made implementing the full lesson difficult. Facilitators needed more time to cover and apply the material. Future implementations of the TSG will need to address scheduling issues on a case-by-case basis.

Sites with lower fidelity scores were in districts that did not use a core reading series (e.g., Virginia). 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 TSG session is to provide time for teachers to plan an upcoming lesson collaboratively. In districts where no core reading series was required, teachers did not have a common lesson that they could discuss. Implementing collaborative planning was difficult as teachers did not follow a sequence of prescribed lessons or couldn't identify specific lesson content.

Another problem we noted was that, in one of the sites, the TSG sessions sometimes became venting grounds for teacher frustrations. Some of our facilitators were skilled in diffusing the situation and getting the participants back on track; others were less skilled and this resulted in not having sufficient time to complete the TSG lesson plans. Specific training and monitoring of facilitators is an issue that needs to be addressed in future research.

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, where the district mandated the use of a core reading series, and where the facilitators were more skilled at keeping the participants focused on the TSG activities. Not surprisingly, fidelity scores were lower in districts from Pennsylvania (75%) and Virginia (69%) than in the California district (95%), where both the facilitators were quite experienced and played a role in developing the curriculum. The Virginia district facilitator was seasoned but the lack of a core reading series in the district created problems in that there was no common set of lessons for teachers to work on. Working together on means for using the research in guided reading was difficult because guided reading typically is very loosely structured. Pennsylvania had the two least experienced facilitators and also the most scheduling problems, with schools only allocating 30 min per week for the TSG sessions.

Districts with lower fidelity scores also did not show consistent effects in both comprehension and vocabulary teaching practices (Pennsylvania: comprehension *d* = 1.12, vocabulary *d* = .34; Virginia = comprehension *d* = .62, vocabulary *d* = 1.13; California: comprehension *d* = 1.33; vocabulary *d* = .80), and the reason for these inconsistencies is not clear.

*Difficulty in implementing Walk through the Research.* An essential assumption of this activity during the TSG program is that teacher will read and come prepared for the sessions. However, broad implementation of this goal remained elusive. Teachers often came to the sessions without reading the required research material. Consequently, the facilitators had to summarize the main tenets of the research material. Having to recap the article or chapter is less than ideal and creating incentives for this remains an issue. Another option is to have teachers read the selections during the sessions to insure that they have the concept that will be highlighted during the session. However, this would take time away from other planned activities during the session. A workable alternative would be to develop a one or two page synopsis of the target research concept. If participants come to the session having not read the material, reading the synopsis during the session would not be time intensive.

### *Limitations of the Study*

Certain limitations of the TSG study are to be noted. The generalizability of the findings is limited to the instructional content focus (i.e., vocabulary and comprehension) of the TSG. Another limitation is that given the complexity of the TSG program and the variations in implementation at each school site, it is not clear how these variations could have affected teacher instruction and outcomes. For example, the TSG format allowed teachers to take their discussions in different directions around a central topic, based on their classroom needs, thus raising the possibility of certain non-targeted changes that have not been measured in this study.

In summary, this study demonstrates a good deal of promise for PD models that (a) focus on findings from scientific research, (b) apply to the existing curriculum in a given school, and (c) facilitate collegial interactions with members of grade level teams, such as the TSG program. 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 PD efforts in first grade (and by implication kindergarten) can potentially benefit from a strong vocabulary and comprehension emphasis.

## References

Adams, M. J., Bereiter, C., McKeough, A., Case, R., Roit, M., Hirshberg, J., et al. (2000). *Open Court Reading.* Columbus, OH: SRA/McGraw-Hill.

American Educational Research Association. (2005). *Research points*. *Teaching teachers*: *Professional development to improve student achievement*. (Vol 3, Issue 1) (brochure). Washington, DC: Author.

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

Baker, S. K., Gersten, R., Dimino, J. & Griffiths, R. (2004). The sustained use of research-based instructional practice: A case study of peer-assisted learning strategies in mathematics. *Remedial and Special Education, 25*, 5-24.

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

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: Guilford Press.

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

Bereiter, C., Brown, A., Campione, J., Carruthers, I., Case, R., Hirshberg, J., et al. (2002). *Open court reading*. Columbus, OH: SRA McGraw-Hill.

Berman, P., & McLaughlin, M. W. (1978). Federal programs supporting educational change. Vol. 8: Implementation and sustaining innovations. Santa Monica, CA: R AND Corporation.

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

Block, C. C., Gambrell, L., & Pressley, M. (2002). *Improving comprehension instruction: Rethinking research, theory, and classroom practice*. San Francisco: Jossey-Bass.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More From Social Experiments: Evolving Analytic Approaches* (pp. 115-172). New York: Russell Sage Foundation.

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*, 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).

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3-15.

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.

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

Calkins, L. M. & Mermelstein, L. (2003). Launching the Writing Workshop. Portsmouth, NH: Heinemann.

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

Cohen, D. K. & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*, 294-343.

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

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

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*, 613-649.

Dewitz, P., Jones, J., & Leahy, S. (2009). Comprehension strategy instruction in core reading programs. *Reading Research Quarterly, 44*, 102-126.

Dole, J. A. (2003). Professional development in reading comprehension instruction. In A.P. Sweet & C.E. Snow (Eds.), *Rethinking reading comprehension*. New York: The Guilford Press.

Dole, J. A., Duffy, G., Roehler, L. R., & Pearson, P. D. P. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61*, 239–264.

Duffy, G. G. (1993). Rethinking strategy instruction: Four teachers' development and their low achievers' understandings. *Elementary School Journal*, *93*, 231-247.

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). New York: Guilford Press.

DuFour, R. (2004). What is a "Professional Learning Community"? *Educational Leadership*, *61*(8), 6.

Durkin, D. (1978). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly*, *14*, 481-553.

Elmore, R.F. (2002) *Bridging the Gap Between Standards and Achievement: Report on the Imperative for Professional Development in Education.* Washington, D.C.: Albert Shanker Institute.

Englert, C. S., & Tarrant, K. L. (1995). Creating collaborative cultures for educational change. *Remedial and Special Education*, *16*, 325-353.

Foorman, B. R., & Moats, L. C. (2004). Conditions for sustaining research-based practices in early reading instruction. *Remedial and Special Education, 25,* 51-60.

Fuchs, L.S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J.D., & Hamlett, C.L. (2005). The Prevention, Identification, and Cognitive Determinants of Math Difficulty. *Journal of Educational Psychology, 97,* 493-513.

Fuchs, L.S., Fuchs, D., Finelli, R., Courey, S.J., & Hamlett, C.L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. *American Educational Research Journal, 41*, 419-445.

Fullan, M. (2001). *The new meaning of educational change*. New York: Teachers College Press.

Gamse, B.C., Jacob, R.T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First impact study final report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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. (1999). Lost opportunities: Challenges confronting four teachers of English-language learners. *Elementary School Journal*, *100*, 37-56.

Gersten, R., Baker, S., & Griffiths, R. (2003). *The sustained use of research based instructional practice: A case study of the early literacy project*. Instructional research group: Unpublished technical report.

Gersten, R., Baker, S., Haager, D., & Graves, A. (2005). Exploring the role of teacher quality in predicting reading outcomes for first grade English learners: An observational study. *Remedial and Special Education, 26,* 197-206.

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., Chard, D., & Baker, S. (2000). Factors that enhance sustained use of research-based instructional practices: A historical perspective on relevant research. *Journal of Learning Disabilities, 33*, 445-457.

Gersten, R., Darch, C., Davis, G., & George, N. (1991). Apprenticeship and intensive training of consulting teachers: A naturalistic study. *Exceptional Children*, 57, 226-237.

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

Gersten, R., & Woodward, J (1992). The quest to translate research into classroom practice: Strategies for assisting classroom teachers' work with "at risk" students and students with disabilities. In D. Carnine and E. Kameenui (Eds.), *Higher cognitive functioning for all students* (pp. 201-218). Austin, TX: Pro-Ed.

Gersten, R., Woodward, J., & Morvant, M. (1992). Refining the working knowledge of experienced teachers. *Educational Leadership*, *49*, 34-39.

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

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: Teachers College Press.

Griffin, G. A. (1995). Influences of shared decision making on school and classroom activity: Conversations with five teachers. *The Elementary School Journal, 96*, 29-45.

Guskey, T. R. (1997). Research needs to link professional development and student learning. *Journal of Staff Development*, *18*, 36-41.

Guskey, T. R. (2003). What makes professional development effective? *Phi Delta Kappan*, *84*, 748-750.

Haager, D., Heimbichner, C., Dhar, R., Moulton, M. & McMillian, S. (2008*). The California reading first year 6 evaluation report*. Morgan Hill, CA: Educational Data Systems.

Harcourt School Publishers. (2005). *Harcourt Trophies*. Orlando, FL: Harcourt School Publishers.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication 16*, 354-367.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371-406.

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

James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., et al. (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.

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

Kennedy, M. M. (1991). Implications for teaching. In E. A. Ramp & C. S. Pederson (Eds.), *Follow through: Program and policy issues* (pp. 57-71). Washington, D.C.: Office of Education Research and Improvement, U.S. Department of Education.

Kennedy, M. (1998). *Form and substance in inservice teacher education* (Research Monograph No. 13). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Kennedy, M. M. (1999, November). *Form and substance in mathematics and science professional development* (NISE Brief Vol. 3, No. 2). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Klinger, J. K., Vaughn, S., Schumm, J. S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth-grade classrooms. *Elementary School Journal, 99*, 3-21.

Lei, P. W., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools, 44*, 433-439.

Lewis, C., Perry, R., Hurd, J. (2009) Improving mathematics instruction through lesson study: A theoretical model and North American case. *Journal of Mathematics Teacher Education*, *12*, 285-304.

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.

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? The case of lesson study. *Educational Researcher, 35*(3), 3-14.

Lieberman, A., & McLaughlin, M. W. (1992). Networks for educational change: Powerful and problematic. *Phi Delta Kappan*, 73, 673-677.

Little, J. W. (2002). Locating learning in teachers' communities of practice: Opening up problems of analysis in records off everyday work. *Teaching and Teacher Education*, *18*, 917-946.

Logan, K. R., & Stein, S. S. (2001). The research lead teacher model: Helping general education teachers deal with classroom behavior problems. *Teaching Exceptional Children*, *33*, 10-15.

Lortie, D. (1975). *Schoolteacher: A sociological study***.** Chicago: University of Chicago Press.

McLaughlin, M. W. (1994). Strategic sites for teachers' professional development. In P. P. Grimmett & J. Neufeld (Eds.), *Teacher development and the struggle for authenticity:*

*Professional growth and restructuring in the context of change* (pp. 31-51). New York: Teachers College Press.

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.

Murphy, C. (1992). Study groups foster schoolwide learning. *Educational Leadership,* 50 (3) 71-74.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

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

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills, 105*, 997-1014.

Pressley, M. (Ed.). (1998). *Reading Instruction that Works: The case for balanced teaching*. New York: The Guilford Press.

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

Pressley, M., & Wharton-McDonald, R. (1998). The development of literacy, Part 4: The need for increased comprehension instruction in upper-elementary grades. In M. Pressley (Ed.), *Reading Instruction that Works: The case for balanced teaching* (pp. 192-227). New York: The Guilford Press.

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

Raudenbush, S. W., Liu, X.-F., Spybrook, J., Martinez, A., & Congdon, R. (2006).

Optimal Design software for multi-level and longitudinal research (Version 1.77) [Computer software]. Available at http://sitemaker.umich.edu/group-based

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*, 479-530.

Sarason, S. (1972). *The creation of settings and the future societies.* San Francisco, CA: Jossey Bass.

Saunders, B., O'Brien, G., Hasenstab, K., Marcelletti, D., Saldivar, T., & Goldenberg, C. (2001). Getting the most out of school-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.

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

Showers, B., Joyce, B., & Bennett, B. (1987). Synthesis of research on staff development: A framework for future study and state-of-the-art analysis. *Educational Leadership, 45*(3), 77-87.

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

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

Sugai, G. (1983). Making a teacher study group work. *Teacher Education and Special Education, 6*, 173-178.

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

Talbert, J. E., & McLaughlin, M. W. (1994). Teacher professionalism in local school contexts. *American Journal of Education, 102*, 123-152.

Taylor, B., & Pearson, D. (2003). The CIERA school change project: *Using data and study groups to improve classroom reading instruction and increase students' reading achievement*. Paper presented at the International Reading Association, Orlando, FL.

Thompson, S. C., Gregg, L., & Niska, J. M. (2004). Professional learning communities, leadership, and student learning. *Research in Middle Level Education Online*, *28*(1), 1-15.

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

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

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*, 80-91.

Warren-Little, J. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal 19*, 325-340.

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.

Wenger, E. (1998). *Communities of practice: Learning as a social system*. Cambridge, MA: Cambridge University Press.

What Works Clearinghouse. (2008, December*). Procedures and standards handbook (version 2.0)*. Retrieved January 2008, from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1

Wiley, D., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis*, *17*, 355-370.

Wilson, S. M., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. *Review of Research in Education, 24*, 173–209.

Woodcock, R. W. (1997). *Woodcock Diagnostic Reading Battery*. Itasca, IL: Riverside.Wright Group. (1996). *Sunshine reading program*. Bothell, WA:Wright Group/ McGraw-Hill.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional.

Footnote

[i] Due to oversampling

[ii] In some schools, TSG sessions were of 30-minute duration; TSG were held more often to keep the total time constant.

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

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

[v] We explored the use of the three *DIBELS* measures *Letter Naming Fluency (LNF)* (*M* = 38.20, *SD* = 15.03, range – 0-87), *Phonemic Segmentation Fluency (PSF)* (*M* = 26.13; *SD* = 16.74, range = 0-74), and *Oral Reading Fluency (ORF*) (*M* = 9.49, *SD* = 12.28, range = 0-109) as 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 (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.

[vi] We replicated the analysis using MANOVA and obtained results that were similar to those in Table 6.

[vii] We replicated the analysis using MANOVA and obtained results that were similar to those in Table 7.

Table 1

*Teacher Demographic Data*

| | | TSG<br>(*N*) | Control<br>(*N*) | *t* | $\chi^2$ | *df* | *p* |
|---|---|---|---|---|---|---|---|
| Initial Sample | | 40 | 44 | | | | |
| Analytic Sample | | 39 | 42 | | | | |
| State | CA | 24 | 29 | | | | |
| | PA | 10 | 9 | | | | |
| | VA | 5 | 4 | | | | |
| Gender | Male | 3 | 4 | | | | |
| | Female | 36 | 38 | | | | |
| University Training | Bachelors | 39 | 42 | | | | |
| | Masters | 13 | 18 | | .78 | 1 | .378 |
| | Post Masters | $6^{\sim}$ | 14 | | 2.82~ | 1 | .093 |
| Certification | Elementary | 39 | 42 | | | | |
| | Reading Specialist | 0 | 1 | | | | |
| | Administrative | 0 | 1 | | | | |
| | Other | 6 | 8 | | | | |
| Total Number of Years of | | *M (SD)* | *M (SD)* | | | | |
| | Classroom Teaching Experience | 11.35 (9.63) | 9.59 (9.76) | -.80 | | 76 | .424 |
| | Teaching in First Grade | 5.05 (5.70) | 4.39 (6.14) | -.49 | | 76 | .623 |
| | Teaching in Current School | 6.05 (6.69) | 6.66 (7.54) | | | | |

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

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 | 217 | 251 | 107 | 128 | 0 | 7 | 26 | 14 | 6 | 6 | 10 | 30 | 18 | 4 |
| California School District | 126 | 167 | 60 | 83 | 0 | 7 | 20 | 11 | 6 | 1 | 5 | 27 | 18 | 4 |
| Pennsylvania School District | 62 | 57 | 32 | 30 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 1 | 0 | 0 |
| Virginia School District | 29 | 27 | 15 | 15 | 0 | 0 | 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

Table 3

*Pretest Scores on Reading and Reading Related Measures for TSG and Control Schools at Baseline*

| Measure | Group | *N* | *M* | *SD* | *t* | *df* | *p* | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|---|---|
| *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)* | | | | | | | | | |
| *Letter Naming Fluency (LNF)* | TSG | 10 | 36.45 | 5.20 | 0.573 | 17 | 0.574 | -4.037 | 7.045 |
| | Control | 9 | 37.96 | 6.24 | | | | | |
| *Phonemic Segmentation Fluency (PSF)* | TSG | 10 | 25.69 | 6.97 | -0.733 | 17 | 0.474 | -9.874 | 4.782 |
| | Control | 9 | 23.15 | 8.17 | | | | | |
| *Oral Reading Fluency (ORF)* | TSG | 10 | 9.03 | 3.06 | 0.392 | 17 | 0.700 | -2.908 | 4.233 |
| | Control | 9 | 9.69 | 4.28 | | | | | |
| *Woodcock Diagnostic Reading Battery (WDRB)* | | | | | | | | | |
| *Reading Vocabulary* | TSG | 10 | 438.94 | 7.03 | -0.178 | 17 | 0.861 | -7.803 | 6.590 |
| | Control | 9 | 438.33 | 7.84 | | | | | |
| *Passage Comprehension* | TSG | 10 | 412.78 | 13.64 | 0.036 | 17 | 0.972 | -14.134 | 14.625 |
| | Control | 9 | 413.02 | 16.07 | | | | | |
| *Oral Vocabulary* | TSG | 10 | 452.84 | 7.18 | 0.067 | 17 | 0.947 | -7.950 | 8.475 |
| | Control | 9 | 453.11 | 9.72 | | | | | |

Table 4

*Professional Development Activities of TSG and Control Teachers*

| Attended Professional Development Activities in | TSG[a]<br>(*N*) | Control[b]<br>(*N*) | $\chi 2$<br>(*df*=1) |
|---|---|---|---|
| Comprehension Strategies | 30 | 24 | 3.56 |
| Vocabulary Instruction | 29 | 23 | 3.38 |
| Phonemic Awareness | 19 | 25 | .95 |
| Decoding & Phonics | 18 | 24 | .98 |
| Fluency | 24 | 28 | .23 |
| Differentiating Instruction | 23 | 24 | .28 |
| Lesson Study-Phonemic Awareness | 20 | 17 | .95 |
| Lesson Study- Decoding & Phonics | 24 | 20 | 1.58 |
| Lesson Study- Fluency | 24 | 28 | .23 |
| Lesson Study- Comprehension Strategies | 13 | 18 | .78 |
| Lesson Study- Vocabulary Instruction | 13 | 16 | .20 |
| Intervention Strategies | 20 | 32 | 5.46 |
| Assessment | 32 | 32 | 0.42 |
| Data Driven Instruction | 30 | 25 | 2.81 |
| Structured English Emersion Techniques & Strategies | 24 | 10 | 11.82* |
| Purposeful Independent Work Time Activities | 22 | 30 | 1.98 |

*Note*. The significance level of .05 was corrected for multiple comparisons using the Bonferroni method (.05/16=.003).

[a]Total number of teachers in TSG = 39. [b]Total number of teachers in Control = 42

* *p* <.05

Table 5

*Power to Detect Effects*

| Teacher Outcomes[a] | MDES Calculated Prior to the Study[c] | MDES Calculated Post Hoc[e] | Observed ES | Actual power |
|---|---|---|---|---|
| Reading Comprehension | 0.65 | 0.76 | 0.86 | 89% |
| Vocabulary | | | 0.58 | 57% |
| **Student Outcomes[b]** | **MDES Calculated Prior to the Study[d]** | **MDES Calculated Post Hoc[f]** | **Observed ES** | **Actual power** |
| WDRB Oral Vocabulary | 0.35 | 0.57 | 0.44 | 58% |
| Reading Vocabulary | | | 0.21 | 18% |
| Passage Comprehension | | | 0.13 | 9% |

Note: MDES (minimally detectable effect size) for all power analyses was based on 80% power using a two-tailed test with alpha set at .05.

[a] Measured by *RCV Observation Measure.* [b] measured by *Woodcock Diagnostic Reading Battery***.** [c] Power Assumptions:19 schools and 95 teachers; alpha ($\alpha$) = 0.05; intra-class correlation *($\rho$)* =.05; variance in posttest explained by pretest covariate *($R^2$)* = 0 (no covariate). [d] Power Assumptions:19 schools and 380 students; $\alpha$=0.05; $\rho$=.05; $R^2$=0.65. [e] Power Assumptions: 19 schools and 81 teachers; $\alpha$=0.05; $\rho$=0.08; $R^2$=0 (no covariate). [f] Power Assumptions:19 schools and 468 students; $\alpha$=0.05; $\rho$=0.22; $R^2$=0.39.

*Table 6*

*Teacher Outcomes*

*Estimated Treatment Effects on RCV Observation Measure*

| | Reading Comprehension | | | Vocabulary | | |
|---|---|---|---|---|---|---|
| Measures | Coefficient | se | t ratio | Coefficient | se | t ratio |
| Fixed Effect | | | | | | |
| Intercept, $\gamma_{00}$ | -0.40 | 0.18 | -2.27* | -0.28 | 0.15 | -1.90 |
| Teacher Study Group, $\gamma_{01}$ | 0.86 | 0.25 | 3.43** | 0.58 | 0.21 | 2.74** |
| Random Effect | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* |
| Between-school variance, $\mu_{0i}$ | 0.13 | 2.53 ~ | 17 | 0.00 | 0.00 | 17 |
| Within-school variance, $\varepsilon_{ij}$ | 0.64 | | | 0.90 | | |

*Estimated Treatment Effects on Content Knowledge for Teaching Reading*

| | Reading Comprehension | | | Reading Vocabulary | | |
|---|---|---|---|---|---|---|
| Measures | Coefficient | se | t ratio | Coefficient | se | t ratio |
| Fixed Effect | | | | | | |
| Intercept, $\gamma_{00}$ | -0.19 | 0.20 | -0.93 | -0.42 | 0.22 | -1.93~ |
| Teacher Study Group, $\gamma_{01}$ | 0.32 | 0.28 | 1.11 | 0.73 | 0.30 | 2.42* |
| Random Effect | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* |
| Between-school variance, $\mu_{0i}$ | 0.13 | 1.12 | 17 | 0.23 | 2.65~ | 17 |
| Within-school variance, $\varepsilon_{ij}$ | 0.86 | | | 0.72 | | |

*Estimated Treatment Effects on Quality Professional Development Scale and Teacher-Teacher Trust Scale*

| | Quality of Professional Development Experienced During the School Year | | | Teacher-Teacher Trust | | |
|---|---|---|---|---|---|---|
| Measures | Coefficient | se | t ratio | Coefficient | se | t ratio |
| Fixed Effect | | | | | | |
| Intercept, $\gamma_{00}$ | -.01 | .16 | -.07 | -.12 | .22 | -.55 |
| Pretest Score, $\gamma_{01}$ | .45 | .07 | 6.06** | .30 | .11 | 2.76** |
| Teacher Study Group, $\gamma_{02}$ | .39 | .22 | 1.76~ | .20 | .30 | .65 |
| Random Effect | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* |
| Between-school variance, $\mu_{0i}$ | .16 | 7.65 ** | 17 | .29 | 4.71* | 17 |
| Within-school variance, $\varepsilon_{ij}$ | .28 | | | .55 | | |

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

Table 7

*Student Outcomes*

*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, $\gamma_{00}$ | -0.13 | 0.15 | -0.88 | -0.21 | 0.18 | -1.13 | -0.11 | 0.14 | -0.78 |
| Letter Naming Fluency, $\gamma_{01}$ | 0.43 | 0.04 | 10.49** | 0.29 | 0.04 | 7.03** | 0.45 | 0.04 | 11.41** |
| Teacher Study Group, $\gamma_{02}$ | 0.21 | 0.20 | 1.06 | 0.44 | 0.25 | 1.73~ | 0.13 | 0.19 | 0.67 |
| Random Effect | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* |
| Between-school variance, $\mu_{0j}$ | 0.16 | 57.21** | 16 | 0.27 | 81.54** | 16 | 0.14 | 57.88** | 16 |
| Within-school variance, $\varepsilon_{ij}$ | 0.67 | | | 0.69 | | | 0.61 | | |

*Estimated Treatment Effects on Reading Accuracy and Fluency*

| | *Letter Word Identification* | | | *Word Attack* | | | *Oral Reading Fluency* | | |
|---|---|---|---|---|---|---|---|---|---|
| Measures | Coefficient | se | t ratio | Coefficient | se | t ratio | Coefficient | se | t ratio |
| Fixed Effect | | | | | | | | | |
| Intercept, $\gamma_{00}$ | -0.15 | 0.14 | -1.08 | -0.11 | 0.16 | -0.66 | -0.13 | 0.11 | -1.19 |
| Letter Naming Fluency, $\gamma_{01}$ | 0.49 | 0.04 | 12.49** | 0.39 | 0.04 | 9.61** | 0.54 | 0.04 | 13.77** |
| Teacher Study Group, $\gamma_{02}$ | 0.21 | 0.19 | 1.09 | 0.13 | 0.23 | 0.57 | 0.23 | 0.16 | 1.47 |
| Random Effect | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* | Variance Component | $\chi^2$ | *df* |
| Between-school variance, $\mu_{0j}$ | 0.14 | 63.75** | 16 | 0.21 | 85.15** | 16 | 0.08 | 30.88** | 16 |
| Within-school variance, $\varepsilon_{ij}$ | 0.62 | | | 0.64 | | | 0.62 | | |

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

Table 8

*Exploratory Analyses: Correlations Between Teacher Measures and Student Reading Outcomes Controlling for Initial Class Performance on Letter Naming Fluency*

| | Teaching Practice[a] | | Teacher Knowledge[b] | |
|---|---|---|---|---|
| Student Reading Outcomes | Comprehension | Vocabulary | Comprehension | Vocabulary |
| *Oral Reading Fluency* | 0.31** | 0.36** | 0.23* | 0.29* |
| *Letter Word Identification* | 0.24* | 0.30** | 0.34** | 0.31** |
| *Word Attack* | 0.30** | 0.33** | 0.36** | 0.28* |
| *Reading Vocabulary* | 0.24* | 0.26* | 0.22~ | 0.27* |
| *Passage Comprehension* | 0.06 | 0.31** | 0.34** | 0.31** |
| *Oral Vocabulary* | 0.21~ | 0.21~ | 0.41** | 0.49** |

[a] Measured by *RCV Observation Measure*. [b] Measured by *Content Knowledge for Teaching Reading*

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

Table 9

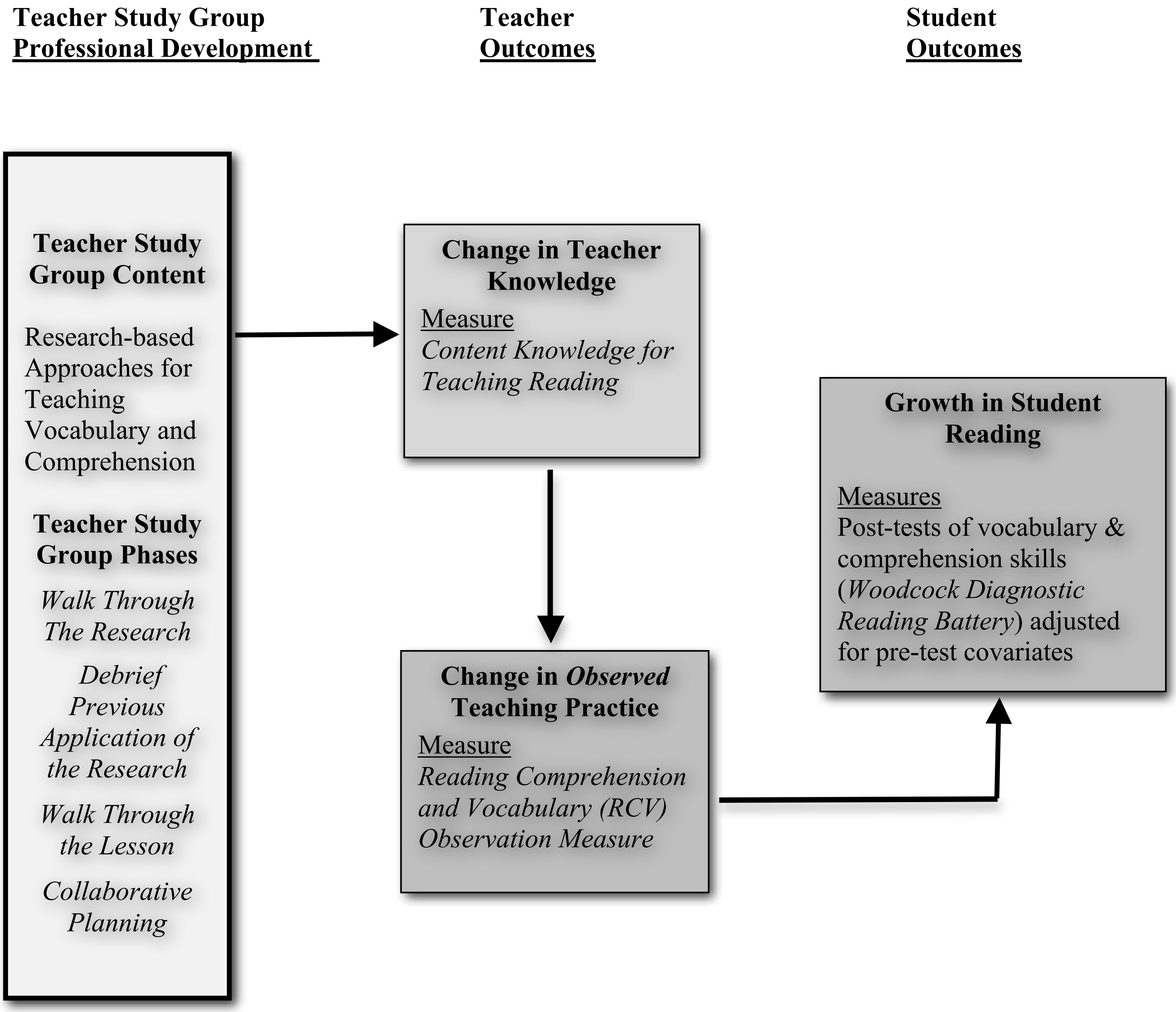
*Exploratory Analyses of Site-Specific Effect Sizes for the RCV Teacher Observationl Measure*

| | *Comprehension* | | | | | *Vocabulary* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *TSG* | | *Control* | | | *TSG* | | *Control* | |
| Site | Effect Size | Number of Teachers | *M*<br>*(SD)* | Number of Teachers | *M*<br>*(SD)* | Effect Size | Number of Teachers | *M*<br>*(SD)* | Number of Teachers | *M*<br>*(SD)* |
| CA | 1.33 | 24 | 1.98<br>(0.77) | 29 | 1.11<br>(0.54) | 0.80 | 24 | 2.81<br>(1.47) | 29 | 1.94<br>(0.70) |
| PA | 1.12 | 10 | 1.25<br>(0.50) | 9 | 0.73<br>(0.43) | 0.34 | 10 | 1.51<br>(0.59) | 9 | 1.28<br>(0.77) |
| VA | 0.62 | 5 | 2.28<br>(0.68) | 4 | 1.93<br>(0.45) | 1.13 | 5 | 2.22<br>(0.60) | 4 | 1.50<br>(0.67) |

*Note*. Effect size = the difference between the posttest means divided by the pooled posttest standard deviation.

Figure Caption

*Figure 1*. Hypothesized Causal Pathways


Teacher Study Group Professional Development
Teacher Outcomes
Student Outcomes
Teacher Study Group Content
Research-based Approaches for Teaching Vocabulary and Comprehension
Teacher Study Group Phases
Walk Through The Research
Debrief Previous Application of the Research
Walk Through the Lesson
Collaborative Planning
Change in Teacher Knowledge
Measure
Content Knowledge for Teaching Reading
Change in Observed Teaching Practice
Measure
Reading Comprehension and Vocabulary (RCV) Observation Measure
Growth in Student Reading
Measures
Post-tests of vocabulary & comprehension skills (Woodcock Diagnostic Reading Battery) adjusted for pre-test covariates

## Appendix A

Sample Items from the *RCV* Observation Measure

| **Comprehension: Sample Items** | | | |
|---|---|---|---|
| ***A. Explicitness of Instruction*** | **Tally** | **Total**<br>**(**Max 15) | **Notes** |
| During or after reading, teacher | | | |
| 1. Models the use of following (includes think-alouds) | | | |
| a. Text cues to interpret text: pictures, sub-headings, captions, graphics | | | |
| b. Visualize events, clarify, re-read. | | | |
| c. Evaluate predictions | | | |
| d. Generate questions about text | | | |
| e. Make text-to-text connections | | | |
| f. Make inferences, summarize/find main ideas- theme, character | | | |
| g. Retell, sequencing – what's happening, what happened first | | | |
| h. Story grammar elements - except for theme, character analysis | N Y | | |
| i. Compare-contrast or cause-effect text structure | N Y | | |
| 2. Reiterates or reinforces concepts that highlight the meaning of text. | | | |
| ***B. Student Practice: Teacher*** | | | |
| 1. Asks students to answer literal recall questions from the text. | | | |
| 2. Asks students questions requiring inferences based on text. | | | |
| 3. Asks students to justify or elaborate their responses. | | | |
| ***C. Corrective feedback: Teacher*** | | | |
| 1. Communicates clearly what student/s did correctly about the strategy. | | | |
| 2. Reinstructs when student makes a mistake by encouraging child to try again or reminding student about comprehension strategy. | | | |
| **Vocabulary: Sample Items** | | | |
| ***A. Explicit Instruction: Teacher*** | | | |
| 1. Provides an explanation, a definition, and/or an example. | | | |
| 2. Elaborates using multiple examples. | | | |
| 3. Uses visuals, gestures, or demonstrations to teach word meaning. | | | |
| ***B. Corrective feedback: Teacher*** | | | |
| 1. Pinpoints the definition further by incorporating ideas from students' responses, examples, and experiences. | | | |

| **Post Observation Component: Sample Items** | | | | | |
|---|---|---|---|---|---|
| 1. Based on your overall judgment, how would you rate the quality of each domain you observed | | | | | |
| | Not observed | Minimal/Erratic | Partially Effective | Good | Excellent |
| Comprehension | | | | | |
| Vocabulary | | | | | |

| 2. How would you rate student engagement today? | | | |
|---|---|---|---|
| | Few students seem engaged | Many students seem engaged | Most students are engaged |
| Students are engaged during the first 45 minutes of the reading block | 1 | 2 | 3 |
| Students are engaged during the remainder of the reading block | 1 | 2 | 3 |