UNIVERSITY OF CALIFORNIA
RIVERSIDE

The Impact of Performance Feedback on Treatment Integrity and Outcomes for a Class-Wide Peer Tutoring Reading Intervention

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Education

by

Yiwen Zhu

June 2014

Dissertation Committee:
Dr. Michael L. Vanderwood, Chairperson
Dr. Sara Castro-Olivo
Dr. Cathleen Geraghty

The Dissertation of Yiwen Zhu is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

The Impact of Performance Feedback on Treatment Integrity and Outcomes for a Class-Wide Peer Tutoring Reading Intervention

by

Yiwen Zhu

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, June 2014
Dr. Michael L. Vanderwood, Chairperson

To improve reading proficiency, evidence-based interventions need to be implemented with integrity in schools. Using a multiple baseline single case design, this study examined the impact of performance feedback from consultants on treatment integrity and student outcomes for Peer Assisted Learning Strategies, an evidence-based, standard protocol, class-wide peer tutoring intervention. Participants were four grade 2 to 5 general education teachers and the students in their classes. Results showed a functional relationship between performance feedback and treatment integrity, including higher treatment integrity of core intervention components. Reading growth rates appeared to have increased, but changes were not statistically significant. Overall, teachers rated the performance feedback procedure and PALS positively. Limitations of the study, implications for practice, and directions for future research are discussed.

Table of Contents

List of Figures

List of Tables

Introduction

Reading is recognized as a skill students need in order to be successful in school and in society. However, the National Center for Education Statistics (2012) found that in 2011, 66% of fourth grade students scored below Proficient in reading on the National Assessment of Education Progress (NAEP). In fact, 33% of fourth graders assessed scored below Basic. These statistics show that a substantial percentage of students need additional support to acquire a level of reading that is considered Proficient (National Center for Education Statistics, 2012).

The growing diversity in the student population has presented additional challenges in reading instruction. Specifically, the number of English Language Learners (ELLs) has increased substantially. In 2010, approximately 4.7 million students (10%) were ELLs, compared to 3.7 million students (8%) in 2001 (National Center for Education Statistics, 2012). ELLs are at greater risk for underperformance in reading compared to native English speakers. In 2011, of 4th grade students who scored above 75th percentile on the NAEP, only 2% were ELLs. Of 4th grade students who scored below 25th percentile, 24% were ELLs, which disproportional to the percentages of ELLs enrolled in public schools (10%). In addition, students from economically and socially disadvantaged backgrounds are particularly at-risk for low performance. Of 4th grade students who scored under 25th percentile in reading, 75% were eligible for free or reduced lunch (National Center for Education Statistics, 2011). Considering these statistics, reading interventions that are effective for diverse groups of students are needed.

With the significant number of students struggling with reading, there has been a substantial amount of research about how to provide effective reading interventions to improve student outcomes. Unfortunately, despite the increased demand for accountability and evidence-based instruction, the 2011 national reading statistics did not improve from 2009 (National Center for Education Statistics, 2012). The persistent percentages of students who are not proficient in reading suggest that even with the availability of evidence-based reading interventions, these interventions are not reaching a sufficient number of students in the population. An inadequate number of schools are adopting these interventions or teachers are not implementing interventions with adequate treatment integrity (Hagermoser Sanetti, Gritter, &, 2011; Kearns et al., 2010).

Treatment integrity, the extent to which an intervention is implemented as designed, is often assumed and not measured (Hagermoser Sanetti & Kratochwill, 2009). When nationally certified school psychologists were surveyed, only 11% responded that they always document treatment integrity when monitoring and evaluating interventions and 33% responded that they never do (Cochrane & Laux, 2008). Interventions that do not include measurement of integrity are at-risk for poor integrity (Hagermoser Sanetti et al., 2011). Furthermore, studies have indicated that teachers implement interventions with low treatment integrity when external support is not provided (e.g., Mortenson & Witt, 1998; Noell et al., 2000; Vadasy, Jenkins, Antil, Phillips, & Pool, 1997). Poor treatment integrity is a concern, as a large number of studies have found that high treatment integrity is related to better student outcomes (Durlak & Dupre, 2008). Low treatment integrity, on the other hand, can result in minimally positive to even detrimental results

(Noell, 2010). Thus, in order to increase and improve implementation of evidence-based interventions in practice, further research is needed on methods to support treatment integrity that are feasible for schools.

In order for a new intervention to be implemented effectively and with integrity, teacher behavior change may be necessary. Consequently, researchers have conceptualized implementation as an adult behavior change process (Noell et al., 2005; Sanetti, Kratochwill, & Long, 2013). School-based consultation is an indirect method of service delivery in which a consultant works with a consultee to change his or her behavior and improve student outcomes (Kratochwill & Bergan, 1990). Performance feedback, in which a consultant monitors implementation and provides feedback to teachers, is a method of consultation that has been found to improve treatment integrity across diverse teachers, students, behaviors, and interventions (Noell et al., 2005). Performance feedback targets components in an empirically supported model of behavior change, the Health Action Process Approach (HAPA), which has been applied to school-based practice. In the HAPA model, self-efficacy (the perceived capability to successfully implement an intervention and affect student learning), outcome expectations, and perception of a problem contribute to behavior intention, which then leads to planning, initiation and maintenance of a new behavior. In this model, self-efficacy plays a role in both behavior intention and actual behavior change (Sanetti et al., 2013). Positive and corrective feedback on implementation and monitoring of student progress promotes teachers' self-efficacy throughout implementation and helps maintain the desired behaviors. In summary, a model of the relationships between performance

feedback, adult behavior change, treatment integrity, and student outcomes would consist of a horizontal progression of each component; performance feedback aims to produce adult behavior change, adult behavior change leads to improved treatment integrity, and higher treatment integrity in turn leads to positive student outcomes. See Figure 1 for a logic model.

In a recent meta-analysis by Solomon and colleagues (2012) of studies that examined the impact of performance feedback on teacher and student behaviors, the majority of studies focused on use of positive reinforcement, implementing behavior plans, and increasing student task completion. Few studies have examined performance feedback for an evidence-based reading intervention. More specifically, performance feedback has not been examined for an evidence-based, class-wide, standard protocol intervention (Solomon et al., 2012). In the standard protocol approach, the same evidence-based intervention and same set of directions is provided for all students with similar skill deficits. The standard protocol approach has a number of advantages over the problem-solving model, in which interventions are designed for each individual. The standard protocol approach allows one evidence-based intervention to reach a large number of students, better promotes consistent implementation across students, and facilitates monitoring of treatment integrity across staff and students. In addition, studies that used the standard protocol approach have demonstrated stronger evidence for improved academic outcomes than studies that have used the problem-solving approach (Fuchs, Mock, Morgan, & Young, 2003). Considering the substantial percentages of students requiring reading interventions and the need to monitor and improve

intervention implementation, studies on performance feedback should target implementation of an evidence-based, class-wide, standard protocol reading intervention. The following is a review of issues in treatment integrity, studies on performance feedback, and research on an evidence-based, standard protocol, class-wide reading intervention.

Treatment Integrity

Treatment integrity terms. Noell (2008) defined three terms to aid in the discussion of treatment integrity in research and practice. Treatment integrity (TI) refers to how accurately the independent variable is implemented in experimental studies. The term intervention plan implementation (IPI) refers to the extent an intervention is implemented as planned in school-based practice. A third term is consultation procedural integrity (CPI), which pertains to the implementation of consultation procedures. Although Noell (2008) differentiates types of treatment integrity using these terms, they are not used consistently in the literature.

Conceptualization of treatment integrity. Treatment integrity is best modeled as a multidimensional construct. Most proposed models of treatment integrity include content (what intervention steps were implemented), quality (how well key components were implemented), quantity, and process of delivery. Power and colleagues (2005) recommend assessing five dimensions of treatment integrity as proposed by Dane and Schneider (1998), adherence, exposure, quality, program differentiation, and participant responsiveness. They further recommend using two measures proposed by Gresham (1989), component integrity and daily integrity for all components combined (Power et

al., 2005). Similarly, eight aspects of implementation according to Durlak and Dupre (2008) are fidelity, dosage, quality, program uniqueness, participant responsiveness, monitoring of comparison conditions, participation rate, and adaptation. Of these eight aspects, the two most studied are fidelity and dosage (Durlak & Dupre, 2008). Although there is not yet a consensus on the definition of treatment integrity, a working definition is "the extent to which essential intervention components are delivered in a comprehensive and consistent manner by a interventionist trained to deliver the intervention" (Hagermoser Sanetti & Kratochwill, 2009, p. 448). This definition assumes there are essential parts of an intervention and stresses that the interventionist should be trained.

Measurement of treatment integrity. Methods to assess integrity include direct observation, self-reports, and permanent products. One inconsistency between measures is the extent of integrity reported by direct observation is typically lower than that reported by self-reports, the latter of which are subjective and can be biased (Noell, 2010). In the study by Wickstrom, Jones, LaFleur, and Witt (1998), the Baseline and Intervention Record teachers used to record child behavior served as a self-report measure of treatment integrity and resulted in an average integrity estimate of 54%. On the other hand, direct observation found an average of only 4% treatment integrity (Wickstrom et al., 1998). An alternative to direct observation or self-report, permanent products can measure treatment integrity easily and reliably, but they cannot measure integrity for components that do not produce permanent products. Of the three methods,

direct observation is the most objective and comprehensive, but also requires more time and resources to collect (Noell, 2010).

Another concern in assessing treatment plan implementation is how detailed each step should be. For example, a checklist with too many detailed steps is not practical and does not make evident the important components of the intervention. In the literature, the primary approach is to describe steps at a level in which steps correspond to observable outcomes and have enough specificity and sensitivity to show a relationship between various levels of integrity and outcomes (Noell, 2010). Noell (2010) recommends using such an approach, in which major, distinct, and measurable components are assessed.

Finally, treatment integrity can be a dependent variable or an independent variable. When the impact of consultation on treatment integrity is the research interest, treatment integrity is the dependent variable. When the relationship of treatment integrity to student outcomes is of interest, treatment integrity is the independent variable (Noell & Witt, 1999). Noell and Witt (1999) provided some guidelines for conducting consultation and treatment implementation research. Mainly, the independent variable of consultation must be experimentally manipulated and related to low and high levels of implementation. This can be done with either single case or group designs, though studies examining the impact of consultation on treatment integrity are typically single case multiple baseline designs (e.g., Mortenson & Witt, 1998; Noell, Witt, Gilbertson, Ranier, & Freeland, 1997; Witt, Noell, LaFleur, & Mortenson, 1997).

Extent treatment integrity is measured in research. In studies on treatment integrity across the years, researchers defined independent variables more often than they

measured them. According to Hagermoser and Kratochwill (2008), the former is more ingrained in practice than the latter, especially for behavioral interventions. In 1989, Gresham coined the phrase “consult and hope,” a method in which consultants do not follow-up with teachers to make sure interventions are implemented as prescribed, when Gresham and Kendall (1987) found that no study on consultation prior to 1987 measured treatment integrity (Gresham, 1989, p. 48). In Gresham, Gansle, Noell, Cohen, and Rosenblum’s (1993) review of 181 school based studies of behavioral interventions from 1980 to 1990, 35% of the studies provided an operational definition of the intervention, 15% assessed and reported treatment integrity information, and 10% stated that they monitored treatment integrity but did not provide data.

Hagermoser Sanetti, Gritter, and Dobey (2011) reviewed 223 single-case and group experimental studies published in five major school psychology journals between 1995 and 2007. The authors coded whether independent variables were operationally defined, the extent to which treatment integrity was monitored, level of risk for poor implementation, treatment agent, location, design, and type of outcome. Most of the studies were conducted at public elementary schools by researchers and teachers. Academic and behavioral outcomes were the most common dependent variables measured. Thirty-two percent provided an operational definition of the treatment and 39% provided a reference to another source for a definition. The authors noted that providing a reference to another source is only appropriate if the intervention was implemented to the same extent as in the reference; if there are any differences, the treatment should be defined. Fifty percent reported treatment integrity data, a substantial

increase from 15% in the Gresham et al. (1993) study. Thirteen percent stated that integrity was monitored but did not provide data, similar to in the Gresham et al. (1993) study. Thirty-seven percent did not mention assessment of treatment integrity and these studies were considered to be at-risk for poor implementation. From the reviews over the years, it seems that the percentage of studies reporting treatment integrity data is increasing but still only about 50%. Consequently, efforts to increase treatment integrity monitoring in school psychology intervention research are still needed.

Of the published studies that measured treatment integrity data, most reported 80% treatment integrity or higher (Hagermoser & Kratochwill, 2008). The meta-analysis by Hagermoser and colleagues (2011), which included mainly studies of academic and behavioral intervention, found an average treatment integrity level of 94%. This suggests that documenting treatment integrity may be a first step in ensuring high integrity. It is also possible that intervention studies that had low treatment integrity did not find positive effects and were not published, increasing the amount of studies that were done with no measurement of integrity or poor integrity (Hagermoser & Kratochwill, 2008).

Extent treatment integrity is measured in practice. Greater emphasis on monitoring treatment integrity in practice is necessary considering the extent it is currently done. Based on a survey given to nationally certified school psychologists by Cochrane and Laux (2008), monitoring of treatment integrity in school-based practice is lacking. While almost 98% of respondents agreed that monitoring treatment integrity is important, 11% said they always do so, 42% said they sometimes do, and 33% admitted that they never do. When asked whether their school's problem solving team measured

treatment integrity, only 2% responded yes. Reasons for why treatment integrity was not monitored included lack of time, lack of administrative support, lack of understanding and willingness from school staff to measure it, and fear that monitoring of integrity would be perceived negatively and as intrusive by teachers (Cochrane & Laux, 2008). As Noell and Gansle (2006) said, "within the culture of schools, it is frequently more comfortable and acceptable to measure student behavior than educator behavior" (p 36). Thus, the drive towards consistent monitoring of treatment integrity should extent to practitioners.

Importance of Measuring and Maintaining Treatment Integrity

Recent focus on getting students to reach standards has increased demand for evidence-based programs (Kearns et al., 2010). There has been greater emphasis placed on evidence-based practice by IDEA 2004 and professional organizations such as the National Association of School Psychologists (Hagermoser Sanetti & Kratochwill, 2009). For an intervention to be considered evidence-based, it must be accurately represent the intervention for which the evidence supports.

Focus on treatment integrity has also increased with the growing use of Response to Intervention (RTI) for special education eligibility determination (Hagermoser Sanetti & Kratochwill, 2009). Treatment integrity is fundamental to a primary goal of the RTI approach to service delivery, which is to provide effective interventions and make decisions based on response to intervention. One concern in RTI is what level of treatment fidelity is sufficient to determine if the student did or did not respond to the intervention that was prescribed (Noell & Gansle, 2006). In evaluating interventions in

research and practice, Durlak and Dupre (2008) emphasized that interpretation of outcomes of an intervention, whether it be positive or negative, can be defensible only if how the independent variable was delivered is known. One cannot say that a program did not work if the program was not implemented as designed. One cannot say that a program worked if it was implemented differently as specified (Durlak & Dupre, 2008).

Treatment integrity contributes to the strength of an intervention. In multi-tiered service delivery models, a common practice to increase the strength of intervention is to increase intensity (Wanzek &Vaughn, 2007). Tier 2 interventions are supplemental to tier 1 interventions and of greater intensity in frequency and duration. Intervention is further intensified at the tier 3 level for students who do not respond to tier 2 interventions (Vaughn, Denton, & Fletcher, 2010). In critique of this practice, Greenwood stated that attempts to increase intervention strength should focus more on quality of intervention components than on intensity of intervention. Hence, Greenwood suggested that increasing intervention treatment fidelity is an important method for increasing likelihood of intervention success that is often overlooked (Greenwood, 2009).

This leads to the main reason to monitor treatment integrity; high treatment integrity is related to positive outcomes (Durlack & Dupre, 2008; Gresham et al., 1993). In Gresham et al.'s (1993) meta-analysis, significant positive relationships were found between percent treatment integrity and intervention effect sizes (e.g., Cohen's d and percent of non-overlapping data). Durlack and Dupre (2008) reviewed five meta-analyses and 59 additional studies focused on interventions conducted in schools and communities by non-researchers. One of the meta-analyses found that programs that monitored

implementation achieved an average effect size of .18 compared to an average effect size of .06 for programs that did not (DuBois, Holloway, Valentine, & Cooper, 2002). Another meta-analysis of 221 school-based behavioral programs found that implementation was the main variable related to outcomes (Wilson, Lipsey, & Derzon, 2003). Of the additional 59 studies reviewed, 76% demonstrated significant positive relationships between treatment integrity and the majority of outcomes measured. In 14% of the studies, there was too little variation in treatment integrity among groups to detect a relationship between treatment integrity and outcomes. Thus, only 10% of the studies did not find a positive relationship between treatment integrity and outcomes. Overall, the literature supports that higher implementation integrity is related to better student outcomes (Durlack & Dupre, 2008).

Factors Affecting Treatment Plan Implementation

The research – practice gap. A main concern for the application of research to school settings is whether interventions found effective in research studies become implemented with high integrity in schools. The first step in transferring research to practice is to gain practitioner interest and attempts to use an intervention. Unavoidably, initial implementation of both non-essential and essential components of the intervention will vary, so the next step is to influence practitioners to implement the intervention with high treatment integrity (Vadasy et al., 1997). As Greenwood (2009) noted, the rewards of research efforts and positive findings are not truly realized until an intervention is successfully brought to scale in practice.

Kearns and colleagues (2010) stated that the research to practice gap is largely due do a long time lack of demand for evidence-based practice. Although greater emphasis on evidence-based practice has emerged, there are barriers that still perpetuate the gap. First, research has traditionally been driven by theory rather than by the problems in schools. Researchers seek to isolate variables of interest and to be able to generalize results, often reducing the importance of the contextual factors (Fuchs & Fuchs, 2001). In addition, researchers and school practitioners work in separate communities and opportunities for collaboration are limited. Researchers converse mainly with other researchers and teachers interact mainly with other teachers (Greenwood & Abbott, 2001). The responsibility for the gap lies with both teachers and researchers. While teachers' knowledge and attitudes affect implementation, it is also argued that researchers should be more aware of the realities of school environments and the needs of teachers. Overall, the consensus is there should be openness and collaboration between researchers and school educators. Finally, some other reasons for poor fidelity in practice are outcomes of implementation are not quickly observable, change is by nature difficult, and change in practice does not directly benefit the teacher but rather the students (Vaughn et al., 2000).

Fuchs and Fuchs (1998) developed and tested a model to bridge the gap between researchers and practitioners, called Project PROMISE (Practitioners and Researchers Orchestrating Model Innovations to Strengthen Education). In this model, researchers work with teachers to identify their needs and incorporate those concerns into the design of an intervention. Next, the intervention is tested using randomized controlled designs.

Finally, the intervention is brought to scale. In fact, this was the process through which Kindergarten PALS was developed. Researchers and teachers met weekly for six months to collaboratively develop the initial version of K-PALS. It was not until the next year that K-PALS was tested in a randomized controlled group study. Results showed that students who participated in K-PALS outperformed control students in phonological awareness and teachers expressed their satisfaction with K-PALS and continued implementing it in following years. From here, the researchers began conducting workshops nationally while continuing to develop the K-PALS program. In summary, Project PROMISE was a successful demonstration of collaboration between teachers and researchers to bridge the research to practice gap (Fuchs & Fuchs, 1998). While researchers agree collaboration is crucial for bridging the research to practice gap, many other factors affect implementation.

Factors affecting implementation. Durlak and Dupre (2008) proposed five categories of factors that affect implementation: treatment program characteristics, service provider characteristics, community factors, school factors, and the support system. One treatment program characteristic that affects treatment integrity is the compatibility of the program with the school's needs and adaptability of the program to meet the school's needs (Durlak & Dupre, 2008). Compatibility with existing instruction is important because teachers are more likely to incorporate a practice is relevant to classroom needs, practical, and easy to implement (Boardman, Arguelles, Vaughn, Hughes, & Klingner, 2005). Gresham (1989) discussed a number of treatment program characteristics that are related to treatment integrity. According to Gresham, interventions

that are more complex, demand too much of teachers' time, and require materials not already available in the classroom are likely to be implemented with lower integrity. In addition, the perceived effectiveness of an intervention, conveyed through student growth data, is related to treatment integrity. Interventions that produce student behavior change more quickly are more likely to be perceived as effective and implemented with integrity.

Provider characteristics include teacher perceptions of need and ability of the intervention to meet the need, teacher self-efficacy, and teacher skills (Durlak & Dupre, 2008). Teacher efficacy includes teacher's belief that they can successfully implement an intervention as well as their belief that they can affect student growth (Ransford, Greenberg, Domitrobich, Small, & Jacobson, 2009; Sanetti et al., 2013). Teachers who have greater self-efficacy are more likely to implement a program successfully, take responsibility for student growth, and set more ambitious goals (Ransford et al., 2009). School or setting factors include the school climate, the extent to which collaboration and communication occurs, and leadership and support from school administrators (Durlak & Dupre, 2008). Lastly, the support system includes training and technical assistance, which should be provided after administrative support, and financial, staff, and time resources have been obtained to implement the intervention (Durlak & Dupre, 2008).

External support. Availability of external support is a main factor that can enhance or hinder sustainability, which is defined as the likelihood an intervention will continue to be implemented with fidelity after supports are removed (Kearns et al., 2010). Kearns and colleagues (2010) proposed four dimensions of external support that may be provided to teachers to impact the likelihood that an intervention will be sustained. They

are level or intensity of support, duration of support, quality of training, and flexibility.

Training and technical assistance are main elements in a support system, and the role of training and technical assistance in intervention delivery is supported by numerous studies (Durlack & Dupre, 2008). While providing teacher training is important, research has found that brief pre-intervention trainings alone are insufficient for supporting implementation (Greenwood & Abbot, 2001; Stein et al., 2008). In the 1970s, professional development often consisted of a one-time training presentation and directions on how to implement the intervention (Gersten, Chard, & Baker, 2000). In the study by Boardman and colleagues (2005), in addition to training, teachers requested classroom demonstrations and asked researchers to observe them and provide feedback. Studies that Durlak and Dupre (2008) reviewed also found that successful trainings included modeling and performance feedback.

Classroom demonstrations and performance feedback represent technical assistance, which should be provided in addition to training. Technical assistance is provided after an intervention has begun, and includes supporting interventionists with problem solving, skill development, and commitment to implementation through monitoring treatment integrity. Early monitoring of treatment integrity followed by additional training can increase treatment integrity to over 80% (Durlak & Dupre, 2008). This is evident in the PALS studies that provided teacher training and implementation assistance. Not only does technical assistance increase treatment integrity, it increases the probability that a practice will be sustained after support is removed (Gersten et al., 2000). However, providing support through researchers is costly and time demanding. It

was estimated that in research projects conducted with schools, the cost to work with each teacher was over $100,000. It was also estimated that more than ten hours per week of researcher time is necessary to produce meaningful change in schools. Such support at a greater scale to reach more teachers would be very expensive and unrealistic (Sindelar & Brownell, 2001).

Balancing flexibility and treatment integrity. Some research suggests that excessive external technical assistance may impede implementation. When evidence-based programs require teachers to strictly follow a manual of prescribed behaviors, it can produce an inflexibility that may be unattractive to teachers who want to modify programs to meet their classroom needs (Kearns et al., 2010). As Kearns and colleagues (2010) noted, flexibility is a dimension of technical support that contributes to sustainability. Understandably, it is difficult to implement interventions exactly as designed when the intervention is complex, and the students, their needs, and the environments are heterogeneous (Noell, 2010). While intervention adaption using clinical expertise may be appropriate, "interventionist drift", the "unplanned, gradual altering of the implementation" is not desirable (Durlak & Dupre, 2008, p. 452). This occurrence is the primary one that monitoring intervention integrity attempts to prevent.

Rigid adherence may sometimes be less effective than allowance of some flexibility and adaptation (Durlack & Dupre, 2008). Using data from the Stein and colleagues (2008) study on Kindergarten PALS, Kearns and colleagues (2010) found that even though more external support was correlated with higher fidelity, participation in the group with the highest level of external support was negatively correlated with

likelihood of sustaining implementation after supports were removed. The authors posited a number of possible reasons for this finding. The presence of a helper may have made the intervention seem more inflexible and reduced teachers' willingness to sustain it. Also, teachers could have become dependent on the helper.

Since flexibility is considered a factor contributing to implementation, fidelity and adaptation can and often occur together in practice. It is suggested that fidelity should be maintained for core intervention components but flexibility can be allowed for less crucial parts. Adaptation may be a reason why treatment fidelity rarely approaches 100% and 80% to 90% is considered high (Durlak & Dupre, 2008). In the studies that found PALS to be effective, fidelity scores were 80% to 90% but sometimes implementation differed from as instructed in the manual. In the Fuchs et al. (1997) study, seven training lessons were conducted rather than all 12 that are in the current manual. In the Mathes and Babyak (2001) study, the order in which components were implemented was changed. Fortunately, 100% integrity is not necessary to produce positive outcomes, as many studies found positive outcomes with less than perfect implementation (Noell, 2010). For example, Durlak and Dupre (2008) found that few studies in their review achieved fidelity above 80%, but many interventions with around 60% fidelity produced positive results. However, studies also show that low levels of implementation undermine intervention success (Noell, 2010).

Greenwood (2009) proposed that research should compare intent-to-treat effect sizes and on-protocol effect sizes. Intent-to treat effects are estimates of effects when an intervention is not implemented with high fidelity. On-protocol effects refer to those

when intervention is implemented with high fidelity. However, a problem with this comparison is choosing cutoffs for high fidelity. As discussed previously, it is unknown how much treatment integrity is good enough (Greenwood, 2009). Although pre-specifying what level of treatment integrity is sufficient is difficult in research studies, in school-based practice, implementation goals can be set depending on student outcomes as demonstrated by progress monitoring. For example, if students are improving substantially though integrity is 70%, that 70% may be appropriate and high enough for these students (Noell, 2010). In order to determine what level of integrity is sufficient for producing positive outcomes, treatment integrity needs to be measured along with student outcomes.

Use of Consultation to Increase Treatment Integrity

Since treatment integrity is related to positive outcomes and there are treatment program, service provider, and support system related barriers to treatment integrity, there is a need for an effective and resource efficient method to reduce barriers to implementation. School-based consultation has been used to increase treatment integrity (Noell et al., 2005). School-based consultation is a collaborative, non-hierarchical, indirect method of service delivery in which a consultant works with a consultee to change teacher behavior and improve student outcomes (Kratochwill & Bergan, 1990). The use of consultation services to enhance implementation of interventions assumes that teachers will change their behavior as a result of consultation processes, so increasing treatment integrity is essentially a matter of changing teacher behavior (Noell et al., 2005; Noell & Witt, 1999). While researchers may advocate different approaches to

consultation, they agree that effective consultation should be collaborative. Generally, when there are enough resources and high accountability, moderate consultation may be enough for adequate implementation. On the other hand, when resources are very limited and accountability and support from administrators are low, even intensive consultation may not be able to drive treatment implementation (Noell & Witt, 1999).

Performance Feedback. Currently, performance feedback is the most supported consultation procedure for increasing intervention plan implementation in school settings (Hagermoser & Kratchowill, 2009). Performance feedback originated in organization psychology (Solomon et al., 2012). In school-based practice, performance feedback is a method of consultation in which a consultant monitors implementation and provides feedback to teachers (Noell et al., 2005). Feedback on implementation progress and student progress increases both teachers' self-efficacy and their perceived effectiveness of an intervention, which are two factors related to treatment integrity (Gersten, 2000; Gresham, 1989). Based on the overall research on performance feedback, 5 to 10 minute weekly meetings, in which treatment integrity and student progress data are reviewed and strengths and weakness of implementation are discussed, are enough to improve treatment integrity (Noell, 2010). Performance feedback has been found to be effective for improving implementation across diverse students, teachers, behaviors, and interventions (Noell, 2010). Teachers also positively rated consultants who provided performance feedback (Noell et al., 2005).

Three seminal studies. In a series of seminal studies examining the impact of performance feedback on intervention implementation, treatment integrity was high in

the beginning but quickly dropped. With performance feedback from a consultant, treatment integrity then increased. Performance feedback also resulted in small increases in student performance (Mortenson & Witt, 1998; Noell, Witt, Gilbertson, Ranier, & Freeland, 1997; Witt, Noell, LaFleur, & Mortenson, 1997). The study by Witt and colleagues (1997) examined performance feedback for an intervention using reinforcement to improve classroom academic performance. Teachers and a student in each teacher's class participated in a multiple baseline study with a pre-intervention training, a post-training baseline phase, a performance feedback phase, and a maintenance phase. Treatment integrity was measured using permanent products. During the pre-intervention training, a consultant assisted with in class implementation and gave corrective feedback. During the post-training baseline phase, teachers implemented the intervention without performance feedback. During the performance feedback phase, at the beginning of each day, the consultant discussed teacher treatment integrity data and student performance data from the previous day. The consultant identified which implementation steps were missed and how to improve implementation. During the maintenance phase, feedback was reduced to once per week. Results showed that treatment integrity decreased during the post-training baseline phase and then increased during the performance feedback phase in both level and trend. Furthermore, treatment integrity remained high during the maintenance phase for three of the four teachers. For three of the four students, academic performance increased from baseline to performance feedback. Percentage correct on daily assignments averaged 71% during baseline, 75% during performance feedback, and 81% during maintenance.

Noell and colleagues (1997) replicated the study by Witt et al. (1997), except they provided pre-intervention training that was less intensive and more typical of school-based consultation. Rather than helping the teacher with implementation on the first day of intervention, the consultant only explained the rationale and procedures and provided materials prior to intervention (Noell et al., 1997, Witt et al., 1997). This extent of pre-intervention training is similar to what teachers receive in the present study. Results were similar to that in Witt et al. (1997); treatment integrity of all three teachers decreased during the baseline phase and clearly increased during the performance feedback phase in both level and trend. For two of the three students, academic performance increased during performance feedback and maintenance phases. This study showed that more intensive pre-intervention training does not necessarily improve teacher integrity when performance feedback is used. It also suggests that intensive pre-intervention training is not necessary when performance feedback will be provided during intervention.

Finally, Mortenson and Witt (1998) again replicated the study by Witt et al. (1997) but decreased the use of performance feedback from daily to weekly. This study found that treatment integrity for two of the four teachers decreased during the baseline phase and increased during the performance feedback phase. Change in student academic performance was more variable than in the previous two studies.

Performance feedback for peer tutoring. Using a multiple baseline design, Noell et al. (2000) examined the effect of performance feedback on teacher implementation integrity for a peer tutoring program. Prior to the start of the intervention, school psychology doctoral students described the intervention to the teachers, provided all

materials, and trained the students on the tutoring routine. On the first days of the intervention prior to baseline, consultants helped teachers achieve 100% treatment integrity. Teacher treatment integrity was measured by whether the teacher provided a session to the student and the correct activity, graded student work, and rewarded the student contingent upon performance. Then during baseline, the teachers were instructed to implement the intervention without interaction with consultants. During the performance feedback sessions, consultants met with teachers for three to five minutes each day to discuss treatment integrity data as well as academic progress data of each student. After teachers achieved 100% treatment integrity four days in a row, performance feedback was reduced to every other day. The mean treatment integrity was 41% during baseline, which was largely due to the fact that for half of the days during baseline, the intervention was not implemented (0% integrity). During the performance feedback phase, treatment integrity varied across the five teachers, but the overall mean increased to 87%. The intervention was implemented on 83% of schools days compared to 50% of school days during baseline. Thus, the main finding of this study was that performance feedback was effective in increasing the number and quality of treatment sessions. However, this study did not measure the integrity of students' tutoring behaviors even though students were mainly responsible for implementing the procedure and their responsibilities were described. The authors stated that measuring student treatment integrity and providing feedback to students should be incorporated in future performance feedback studies. Hence, the present study will measure both teacher and student treatment integrity. Lastly, while this study examined the effect of peer tutoring

on student performance, it did not examine the impact of performance feedback on student outcomes.

Greenwood and colleagues (2001) studied the impact of consultation on treatment integrity and student outcomes for a Class-wide Peer Tutoring Learning Management System (CWPT-LMS) in an elementary school. The LMS is a computer software program that helps teachers set up and graph student progress. Their study used single case AB design, weekly pre and post vocabulary and spelling tests, and a 40-item procedural checklist for CWPT-LMS implementation that included teacher and student behaviors. The five classroom teachers who participated in this study initially learned how to implement CWPT using the manual with the help of consultants until they reached 80% integrity. They were also trained on how to use the CWPT-LMS. For the next five to seven weeks, teachers implemented CWPT and used the CWPT-LMS system on their own. Then, a consultant met with each teacher for one hour to analyze student progress and make plans on how to improve the program. Every two weeks for the rest of the school year, the consultants provided teachers with feedback regarding implementation and student progress and discussed further changes to the program. The average monthly treatment integrity score for the entire period of CWPT implementation was high, at 97%. After consultation began, the percentage of students who were "successful" increased from 35% to 58%. Success was defined as a score of 40% or less on the weekly pretest and a score of 80% or more on the weekly post-test, or in other words, an increase of 40% from pre to post test for that week. Although this study found that performance feedback was related to better student outcomes, it was an AB design,

so did not have experimental control. Teachers in this study also had the additional benefit of a data analysis tool in addition to receiving feedback from consultants.

Performance feedback compared to two other methods. A more recent study by Noell and colleagues (2005) used a randomized controlled trial design to compare the impact of performance feedback on treatment integrity to that of two other methods: brief weekly interviews and brief weekly interviews with an emphasis on commitment to the intervention. Commitment was used as a social influence strategy. A behavioral consultation model was initially used with 45 teachers to develop behavioral treatment plans for students. Teachers then participated in one of the three follow-up conditions. In the interview condition, a consultant met with a teacher weekly to ask about the extent of plan implementation and student improvement and gave the teacher a chance to ask questions. In the interview and commitment condition, a discussion including five key points about commitment to intervention was added in the last interview. In the performance feedback condition, a consultant met with the teacher every day to review permanent products, intervention implementation, and student progress data, similar to the performance feedback methods used for reading interventions. Treatment integrity as measured by permanent products during the three weeks of intervention was significantly higher in the performance feedback group than in the other two conditions. In the other two conditions, treatment integrity was not significantly different. The study also found that treatment integrity decreased from week one to two and remained statistically the same at week three, but it appeared that the decrease between week one and two was smallest for the performance feedback condition. Treatment integrity for the group that

received only a weekly interview dropped the most over the three weeks. Student behavior change was also significantly greater for the performance feedback condition and not significantly different for the other two conditions. This study suggests that weekly interviews and encouraging teachers to remain committed to an intervention are not enough to sustain treatment integrity. On the other hand, frequent performance feedback may be able to buffer against the tendency for integrity to drop over time. Furthermore, teachers rated the performance feedback consultation procedure positively, with a mean of 6.5 on a 7-point Likert scale.

Recent meta-analysis. Solomon, Klein, and Politylo (2012) recently published a meta-analysis on single case studies that further supported the effect of performance feedback on treatment integrity. The following criteria were included in selecting the studies. The study used Noell's (2005) definition of performance feedback, which is "monitoring a behavior that is the focus of concern and providing feedback to the individual regarding that behavior" (p. 88). The focus of the study was changing teacher behavior. The study used single case design; this included both studies with three replications of effect such as ABAB and multiple baseline designs and studies with fewer than three replications of effect such as AB designs. The main dimensions coded in the performance feedback studies were type of intervention, setting, and immediacy of feedback. Interventions were behavioral (such as reinforcement schedules and discrete trials) and academic (including peer-tutoring and goal setting). Settings included preschools, elementary, middle, and high schools. Performance feedback delivery ranged

from immediate to delayed by a week, and in some studies it was not reported (Solomon et al., 2012).

Sixty-nine percent of the studies in the meta-analysis reported treatment integrity information. The authors calculated an effect size for each study called ALLISON-MT, which is a regression correlation coefficient (*R*) that models level and trend and adjusts for autocorrelation. The authors found that the average effect of performance feedback was positive and medium for both teacher and student behaviors, though there was a large amount of variation across studies. The average *R* for the effect of performance feedback on teacher behavior was .72, a medium effect. Individual effect sizes ranged from $r = -.27$ to $r = .99$. The authors noted that the medium positive effect was more significant than it appears because a zero effect size meant teacher integrity remained constant, which is often not the case without performance feedback. In fact, consistent with other research, the average slope of integrity during baseline in these studies was -2.83 ($SD = 4.29$), which means integrity tended to decrease without feedback. The effect of performance feedback on student behavior based on 52 student participants was $R = .50$. In individual studies, student effects ranged from $r = -.35$ to $r = .78$. Compared to the effect of performance feedback on integrity in behavioral intervention, the effect in academic interventions was significantly higher. Lastly, the effects of immediate performance feedback, performance feedback within one day, and weekly performance feedback were not significantly different (Solomon et al., 2012).

The above studies showed that performance feedback is effective for increasing treatment integrity. Performance feedback was provided at least once per week based on

objective and measurable observations of teacher integrity and student progress. Furthermore, the studies suggested that neither intensive training nor daily performance feedback is necessary to produce effects (Noell et al., 1997; Solomon et al., 1997). Instead, a pre-intervention training in which procedures are explained and materials are provided coupled with weekly brief performance feedback could increase treatment integrity (Mortenson & Witt, 1998; Noell, 2010). Generally, there were improvements in student performance overall during the performance feedback phases, though that change was less consistent than the change in teacher integrity (Solomon et al., 2013)

Limitations in performance feedback research. Although the existing literature on performance feedback supports that it is effective for increasing treatment integrity and outcomes in schools, previous studies have a number of limitations. Much of the previous research, including the studies conducted by Witt and colleagues (1997), Noell and colleagues (1997), and Mortenson and Witt (1998) that are continuously referenced in newer studies, used multiple baseline single case design but were conducted prior to the establishment of single case design methodological standards (Smith, 2012; Solomon et al., 2012). In the study by Mortenson and Witt (1998), an increase in treatment integrity from baseline to performance feedback phase was observed for only two of the four teachers; according to present standards, a casual relationship between an intervention and outcomes cannot be concluded without three replications of effect (Kratochwill et al., 2012). Second, in the studies above, treatment integrity was measured only by permanent products rather than by direct observation. Validity, reliability, and inter-observer agreement information was not reported for the treatment integrity

checklists. In addition, student progress was not measured using current, psychometrically validated measures of reading outcomes (CBM) (Mortenson & Witt, 1997; Noell et al., 1997; Witt et al., 1997). Regarding data analysis, most studies did not estimate effect sizes, nor report enough actual qualitative data and datasets had to be recreated from graphs (Solomon et al., 2012).

An additional gap to the literature is only one performance feedback study (Noell et al. 2005) examined teacher acceptability of the performance feedback procedure. Teacher acceptability is a measure of the social validity of an intervention, which includes the social significance of goals, social appropriateness of procedures, and importance of intervention effects (Wolf, 1978). Researchers hypothesize that teacher acceptability of an intervention affects treatment integrity (Noell, 2010). In the case of performance feedback, teacher acceptability of performance feedback should affect their likelihood to correct PALS implementation in accordance with the feedback given.

Finally, studies have examined performance feedback targeting the use of positive reinforcement, implementation of behavior plans, increasing student task completion, and implementation of peer tutoring math and reading interventions. However, no published studies have examined the effect of performance feedback on treatment integrity and student outcomes for an evidence-based, class-wide, standard protocol reading intervention (Solomon et al., 2012). The standard protocol approach is research supported and resource efficient; it uses the same evidence based intervention for all students with similar problems. In the standard protocol approach, the same set of instructions is given to groups of students (Fuchs et al., 2003). Thus, a standard protocol intervention would

be a valuable medium through which to examine the effectiveness of performance feedback. Since the effect of performance feedback on teacher and student behavior varied greatly across studies (Solomon et al., 2012), the effectiveness of performance feedback for a standard protocol class-wide reading intervention is difficult to predict.

Standard Protocol, Class-wide Reading Interventions

Delivering Reading Interventions. A multi-tiered model is commonly used in schools to effectively allocate educational resources to students (Tilly III, Niebling, & Rahn-Blakeslee, 2010). At the tier 1 level, interventions are delivered in the classroom by teachers and can extend to all students in general education. Such interventions may address a weakness in tier 1 instruction or be preventative (Tilly III et al., 2010). At the tier 2 level, small group interventions are used to provide additional targeted support for struggling students. However, tier 2 interventions are resource intensive. Effective tier 2 interventions should be delivered in small groups of three to five students for 30 minutes a day and three days a week for at least eight weeks (Vaughn, Denton, and Fletcher, 2010). Thus, in schools with a substantial proportion of struggling readers, tier 1 interventions may be more resource efficient. A group of tier 1 reading interventions that are effective and warrant attention are class-wide peer tutoring interventions (U. S. Department of Education, 2012b).

Class-wide Peer Tutoring. An instructional strategy often used in schools to improve student academic performance is peer tutoring, in which students help each other learn a skill or idea (Thomas, 1993). Peer tutoring is designed to provide differentiated instruction and increase the teacher to student ratio, academic engagement, and

opportunities for feedback to students (Greenwood, Arreaga-Mayer, Utley, Gavin, & Terry, 2001). From a cognitive developmental perspective, peers also promote cognitive development through scaffolding, motivation, and social interaction (Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003). In traditional forms of peer tutoring, tutoring is one-directional. A student who is more skilled from a higher grade tutors a younger student (Greenwood et al, 2001). Another type of peer tutoring is reciprocal peer tutoring. First developed by Fantuzzo and colleagues, reciprocal peer tutoring allows students to maximally benefit from peer tutoring by alternating between being the tutor and the tutee. Pairs are encouraged to work as a team to prompt, check, and assess each other while gaining knowledge of a certain academic area (Fantuzzo, Polite, & Grayson, 1990).

Peer tutoring is often classroom-based. It may be conducted in pairs or with small groups of students in various academic areas (Rohrbeck et al., 2003). Four models of class-wide peer tutoring are Class-wide Peer Tutoring (CWPT), the START tutoring program from Ohio State University, Class-wide Student Tutoring Teams (CSTT), and Peer Assisted Learning Strategies (PALS) (Maheady, Mallette, & Harper, 2006). The oldest program is CWPT, which was developed, tested, and validated in the 1980s (Greenwood, Terry, Arreaga-Mayer, & Finney, 1992). In CWPT, reciprocal peer tutoring is conducted with an entire class of same aged students (Greenwood et al., 2001). Goals in the development of CWPT were it would benefit all students, use materials already available in the classroom, supplement instruction, and require no additional time or work from the teacher (Delquadri, Greenwood, Stretton, & Hall, 1983). Whereas CWPT and PALS group the whole class into dyads, CSTT groups the class into teams of four, and

the START model allows tutoring to be conducted for certain students, in small groups, or in pairs class-wide. In all four models, tutoring is reciprocal, students are training on tutor and tutee roles, teachers actively monitor students, and group contingencies are used (Maheady et al., 2006).

A meta-analysis by Rohrbeck and colleagues (2003) of 81 peer assisted learning intervention studies found an effect size of .59 (*SD* = .90). For reading outcomes, 26 studies averaged a Cohen's effect size of 0.26, which is considered small to moderate (Cohen, 1988). Interestingly, studies that had larger effect sizes tended to have participants who were younger (e.g., grades 1 to 3) and from more diverse backgrounds (e.g., urban and at least 50% minority). This suggests that class-wide peer tutoring may be especially beneficial for the more at-risk students (Rohrbeck et al., 2003). For CWPT interventions, 25 published studies found CWPT to produce greater reading and comprehension outcomes compared to traditional teacher-led instruction. Effect sizes were at least .40 (Greenwood et al., 2001). Of the peer tutoring models, CWPT and PALS have been researched the most and accumulated the most evidence (Maheady et al., 2006).

Peer Assisted Learning Strategies (PALS)

PALS is a research supported, class-wide reciprocal peer tutoring program that targets alphabetics, fluency, and comprehension. PALS was developed by Lynn and Doug Fuchs in 1997. PALS for Kindergarten (K-PALS) and first grade PALS focuses on phonological awareness and phonics while PALS for grades 2 to 6 focuses on fluency and comprehension skills. Each has its own manual with scripted lessons and materials.

The manuals are designed to be explicit, complete, and easy to use. To pair students, students are first ranked by reading performance and split into a higher group and lower group. The first ranked reader from the higher group is paired with the first ranked reader in the lower group. The second ranked reader from the higher group is paired with the second ranked reader in the lower group and so forth. Students in each pair take turns reading and being the peer tutor (McMaster, Fuchs, & Fuchs, 2007).

The PALS program, while empirically supported, encompasses a number of barriers to treatment integrity as stated by Gresham (1989). It is procedurally complex with a large number of specific, required teacher and student behaviors. Student materials need to be photocopied and books need to be acquired. Furthermore, it may take time for teachers to perceive the intervention as effective, as PALs targets fluency and comprehension skills, which may take time to show on general outcomes measures (CBM). Finally, based on previous research on PALS (Vadasy et al., 1997), PALS is difficult for teachers to implement without support and feedback. Thus, PALS is useful medium through which to study the effectiveness of performance feedback on reducing intervention barriers to treatment integrity.

PALS Research. PALS is one of the few tier 1 reading interventions supported by What Works Clearinghouse (U. S. Department of Education, 2012a; U. S. Department of Education, 2012b). WWC is a recognized resource for evaluating interventions that was created in 2002 by the U. S. Department of Education's Institute of Educational Sciences (IES) to review, summarize, and report research evidence in support of educational interventions. Its goal was to make this information reliable and accessible to

educators and researchers in order to improve education practices. First, WWC identifies studies that are relevant and eligible for review, based on criteria including time of publication, methodological design, appropriate age range of participants, and whether the effect of intervention was analyzed based on student outcomes. Then, WWC uses rigorous standards to determine which studies should contribute to the evidence base and only includes studies that meet standards or standards with reservations. Randomized controlled trials with correct methodology and low attrition meet standards while quasi-experimental designs with equivalent experimental and control groups meet standards with reservations (U. S. Department of Education, 2011).

WWC examined research studies on PALS and concluded that PALS has potentially positive effects for alphabetics and comprehension (U. S. Department of Education, 2012a; U. S. Department of Education, 2012b). Of approximately 15 published studies that examined the effectiveness of K-PALS and first grade PALS on early literacy skills and were eligible for review, two met WWC standards (McMaster, Fuchs, Fuchs, & Compton, 2005; Stein et al., 2008) and one met standards with reservations (Mathes & Babyak, 2001) in the 2012 report. Based on these three studies, the evidence in support of K-PALS and first grade PALS was considered medium to large (U. S. Department of Education, 2012a). Of four published studies that examined the effectiveness of PALS for grades 2-6 and were eligible for review, one met standards (Saenz, Fuchs, and Fuchs, 2005), and one met standards with reservations (Fuchs, Fuchs, Mathes, & Simmons, 1997) in the 2012 report. Based on these two studies, the evidence in support of PALS for grades 2 to 6 was small (U. S. Department of Education, 2012b;

U. S. Department of Education, 2010). Although studies on PALS with sound methodology have found positive results, they have limitations in terms of generalizability of results to other schools and students. It is important to know that in the above studies, PALS was implemented with substantial support from researchers and high levels of treatment integrity, which may not be present at other schools (McMaster et al., 2007). The following is an analysis of treatment integrity and outcomes in PALS studies conducted with high implementation support as well as in a study with low implementation support. High support includes daylong training workshops combined with regular in-class implementation assistance whereas low support consists of a brief training without follow-up support (McMaster et al., 2007; Vadasy et al., 1997).

PALS with high implementation support. Mathes and Babyak (2001) examined outcomes for low achieving, average achieving, and high achieving first graders who participated in PALS and PALS with additional mini lessons. Prior to implementation, teachers received the PALS manual, additional daily lessons created by the researchers, and a full-day training. During implementation, researchers provided in class support, and PALS was implemented three days a week, 35 minutes each day, for 14 weeks. Students who participated in first grade PALS with and without additional mini-skills lessons grew more on progress monitoring measures of phonological awareness and oral reading fluency compared to students who had regular reading instruction. They also grew more from pre-test to post-test on the Word Identification and Word Attack subtests of the Woodcock-Johnson Reading Mastery Test – R with effect sizes of .67, .90, and .60 for low, average, and high achieving students respectively.

McMaster, Fuchs, Fuchs, and Compton (2005) compared first grade PALS to one-on-one intervention from an adult for 66 low performing students (non-responders) who had Curriculum-Based Measurement (CBM) scores and growth rates a half standard deviation below average performing students. First, 33 first grade teachers were randomly assigned to a PALS condition, a modified fluency building PALS condition, and a no PALS condition. After teachers in the PALS and modified PALS conditions implemented PALS for seven weeks, 20% of the students who participated in PALS were identified as non-responders. This also means that 80% responded to the intervention. The non-responders were randomly assigned to continue with PALS, modified PALS, or one on one tutoring for another 13 weeks. Results showed that among the PALS, modified PALS, and adult tutoring groups, there were no significant differences in non-responders' post-test scores controlling for pre-test scores, as well as no significant differences on progress monitoring measures. This suggests that PALS may be as beneficial to low performing students as one on one adult intervention. Since PALS requires only one teacher per class for implementation whereas one on one tutoring requires one teacher per student, PALS appears to be the more practical and efficient choice if it can produce outcomes similar to one on one tutoring. Overall, this study supports that when PALS is implemented with high integrity (92% in this study), it is effective for 80% of tier 1 students as well as for low performing students, and may be a better use of resources than one on one adult tutoring. Furthermore, the finding that there were no differences in outcomes between the PALS and PALS with fluency groups and between the PALS and modified PALS groups supports that changes to the original

PALS did not make PALS more effective.

Stein et al. (2008) conducted a study of PALS that included approximately 3,000 kindergarteners from 67 schools in three states. Teachers were randomly assigned to (1) a no PALS group, (2) a PALS with training workshop only group, (3) a PALS with workshop and booster training session group, or (4) a group that had PALS with a workshop, booster, and research assistant as an in class helper two days a week. In each PALS condition, teachers implemented PALS four days a week for 18 weeks. Treatment fidelity averaged 86%. Results found that students in the PALS with booster training group and the PALS with booster training and helper group made greater gains in letter sounds naming compared to students in the no PALS group and the PALS group that received just the initial workshop. Effect sizes for the PALS with booster group and the PALS with booster and helper group were 1.18 and 1.02 respectively. This suggests that a booster training or classroom helper increases the effectiveness of PALS.

In fact, the researchers found that treatment fidelity was significantly different by level of teacher support. The group that received all three types of support had the highest fidelity scores, followed by the group that received the workshop and booster, followed by the workshop only group. The researchers also showed that student gains in the different conditions were mediated by implementation fidelity because when fidelity scores were included in the regression, the effect of treatment conditions became non-significant. In other words, treatment fidelity explained variance in student outcomes (Stein et al., 2008).

Two studies on PALS grades 2 to 6 found positive outcomes for students in

general education, students in special education, and English Language Learners (Fuchs et al, 1997; Saenz et al., 2005). Fuchs and colleagues (1997) examined PALS for students in grades 2 to 6 who were average achieving, low achieving, or qualified as learning disabled. In this study, teachers assigned to the PALS condition attended a full-day training workshop and then conducted PALS three times a week for 15 weeks. During the first seven PALS lessons, teachers received assistance from the researchers. Mean treatment integrity for teacher behaviors was 89% and mean treatment integrity for student behaviors was 86%. In each class, one average achieving, one low achieving, and one learning disabled student was assessed using the Comprehension Reading Assessment Battery (CRAB), which includes an oral reading task, a maze comprehension task, and 10 open ended comprehension questions. Pre and post-test scores demonstrated statistically significant differences in growth between the PALS and no PALS groups. Effect sizes on each CRAB score were .22, .56, and .55 respectively. Another important finding was that there were no significant differences in growth by type of student, supporting that PALS can similarly benefit average, low achieving, and students with learning disabilities. As with the K-PALS and first grade PALS studies, this study supports that PALS conducted with classroom support and high treatment integrity is effective for improving reading outcomes.

Recognizing that participants in previous studies were mainly native English speakers, Saenz, Fuchs, and Fuchs (2005) examined the effectiveness of PALS for English Language Learners (ELLs) in grades 3 to 6. The authors used similar methodology and provided similar levels of teacher support as Fuchs and colleagues

(1997) did, except they included only Spanish speakers, many of whom were struggling in reading. Although the students were in grades 3 to 6, the mean reading grade level was about 3.5. In each participating class of ELL students, researchers identified three low achieving, three average achieving, three high achieving, and two learning disabled students to monitor. Teachers in the PALS condition implemented PALS three times a week for 15 weeks and achieved an average treatment integrity of 94% for both teacher and student behaviors. For the questions correct score on the CRAB, there was a significant main effect of PALS; students participating in PALS grew more from pre-test to post-test compared to the contrast group, with an average effect size of 1.02. However, the CRAB fluency and comprehension scores did not show significant differences in growth between groups. The limited significant findings may have been due to the lack of power in this study, because the unit of analysis was teachers and each condition only had six teachers. Since this is the only study with sound methodology that supports PALS grades 2 to 6 for ELLS, and the extent of evidence is small, additional research on PALS should include more ELL students.

PALS with low implementation support. Since studies that found PALS to be effective provided in-class implementation support, a question arises as to what level of treatment integrity might teachers achieve without in-class support. A study conducted with PALS for grades 2 to 6 found that when teachers received only a 40-minute introduction to PALS and the PALS manual, they implemented the program with poor integrity (Vadasy et al., 1997). The purpose of this investigation was not to examine PALS outcomes, but to examine fidelity of implementation when teachers are offered

low or high support. The low support group was provided with only an introduction to PALS and the manual. The high support group was provided an introduction and the manual and also offered a $50 award for reading the manual, a one time half day substitute to allow teachers to observe PALS, assistance from researchers, and another $50 for completing weekly treatment integrity logs. However, the latter group did not take advantage of the implementation assistance that was offered. Of the 15 teachers in the low support group, four (27%) initiated PALS in their classrooms but only two (13%) completed 15 weeks. Of the 29 teachers in the high support group, six (21%) initiated PALS but just four (14%) implemented it for 15 weeks. For both groups, the weekly progress logs were not completed half the time and were of poor quality. All teachers made substantial changes to the PALS implementation. They combined activities, omitted activities, reordered activities, increased or decreased the amount of text read during activities, modified students' responsibilities, and neglected the points system. For all these changes, teachers reported rational reasons for doing so, highlighting that the difficulty with maintaining treatment integrity is typically not due to negligence, but to a desire to adapt the program to perceived classroom needs. This study provides a realistic depiction of how PALS may be implemented without in-class support.

Limitations in PALS research. The existing research on PALS summarized in the above studies presents several limitations. One limitation is a small number of studies have examined PALS for grades 2 to 6. Only four studies were eligible for review and the evidence in support of PALS for grades 2 to 6 is based on only two studies (U. S. Department of Education, 2012b; U. S. Department of Education, 2010). Second, it is

unclear how effective PALS would be in schools with large percentages of ELL students. All the studies except the one by Saenz et al. (2005) were conducted with primarily native English speakers in regions with relatively small ELL populations. Mathes and Babyak's (2001) study was conducted in a southeastern medium sized school district. Student participants were approximately 50% Caucasian and 50% African American. Fuchs and colleagues' (1997) study was conducted in suburban and urban districts in a southern state. Seventy-eight percent of student participants were Caucasian. The study by McMaster and colleagues (2005) took place in Nashville, Tennessee. Support for PALS for ELLs is based on only one study (Saenz et al., 2005), and WWC considers the extent of evidence to be small (U. S. Department of Education, 2010). Thus, it is important to further examine the effectiveness of PALS for ELLs.

A major limitation of these studies is the amount of implementation support provided in the studies that found positive results. Researchers directed implementation and provided full-day training workshops and in-class support. In McMaster et al. (2005), teachers participated in a full-day training and received visits from support staff two times a week during the first seven weeks and once a week during the next 11 weeks. In Fuchs et al. (1997) and Saenz et al. (2005), teachers first attended a full-day workshop then received assistance from graduate students about once a week. McMaster and colleagues (2007) acknowledged that the trainings and technical support provided largely contributed to the high integrity and effectiveness of PALS in the above studies. In fact, Stein and colleagues (2008) found that treatment integrity differed by level of teacher support and mediated the relationship between level of support and student outcomes

(Stein et al., 2008). Based on these studies, it is unclear whether the magnitude of the positive results found can be expected in schools not participating in effectiveness research studies. As the study by Vadasy and colleagues (1997) demonstrated, teachers conducted PALS with poor integrity without implementation support. Since schools do not typically receive a large extent of implementation support, studies should examine PALS implementation in school settings using more practical methods of providing teacher support, such as performance feedback (McMaster et al., 2007; Vadasy et al., 1997). While studies have found performance feedback to improve treatment integrity and student outcomes in peer tutoring interventions (e.g., Greenwood et al., 2001; Noell, 2000), none have been conducted with PALS.

Finally, although studies on PALS reported treatment integrity, they did not examine treatment integrity for specific components of PALS. Treatment integrity may have been high for core components, or perhaps there were certain core components that were consistently altered across teachers. Vadasy and colleagues (1997) identified 20 key components of PALS and found that certain components were most frequently implemented incorrectly. In all classrooms observed in their study, the points system and student question cards were not used. In the majority of classrooms observed, the higher reader did not read first, students did not sit side by side, and the PALS activities were not timed (Vadasy et al., 1997). As Durlak and Dupre (2008) stated, treatment fidelity is most important for core intervention components. Thus, further research on PALS implementation should examine the extent to which key components of PALS are implemented with fidelity.

Purpose of the Study

The purpose of this study was to examine the impact of performance feedback on treatment integrity and student outcomes for a standard protocol, class-wide reading intervention for grades 2 to 6. A unique characteristic of this study is that it examines the implementation of PALS in an urban school district with large percentages of Spanish speaking ELL, low SES, and at-risk students. A multiple baseline single case design was used to examine the level of treatment integrity achieved when teachers are provided with only an initial training presentation, student materials, and the PALS manual, and subsequently the impact of performance feedback on treatment integrity. The study compared the extent to which key components of PALS are implemented, with and without performance feedback. In addition, the study examined growth rates for ELLs and native English speakers to support the link between performance feedback and student outcomes. Lastly, the study examined teachers' acceptability of PALS and the performance feedback procedure. The research questions were as follows:

1. What is the average level of treatment integrity to which PALS for grades 2 to 6 is implemented when teachers are not provided with performance feedback (e.g., with only a training presentation, student materials, and the PALS manual)?
2. Is there a functional relationship between providing performance feedback to teachers and an increase in the treatment integrity of PALS implementation?
3. To what extent are core components of PALS implemented, without performance feedback and with performance feedback?

4. Does providing performance feedback to teachers on PALS increase growth rates for ELL students and native English speakers?

5. To what extent do teachers perceive performance feedback and PALS as useful?

Method

Experimental Design

A concurrent multiple baseline single case design was used to examine the impact of performance feedback on PALS treatment integrity after teachers had first attempted to reach full implementation independently. Single case designs (SCDs) are experimental designs used to examine the effectiveness of interventions. Rather than assigning participants to experimental and control groups, SCD uses single subjects (e.g., students, teachers, classrooms) and each subject serves as its own control. A multiple baseline design is a series of A-B designs in which implementation of the same intervention begins at different times for each subject. Experimental control and intervention effect are demonstrated through multiple replications of effect (Kratochwill et al., 2010).

Several professional groups (e.g., WWC, APA Division 16, the National Reading Panel) have written sets of SCD research guidelines for intervention research. While the guidelines generally agree on main aspects of design, they have some variations reflecting the specific purposes the guidelines were designed for. Of the prominent SCD research guidelines available, the WWC methodological standards are most detailed and rigorous (Smith, 2012). Therefore, WWC standards for SCD methodology were used to guide design of this study. They are as follows:

1. The independent variable must be systematically manipulated.

2. The dependent variable should be measured systematically and repeatedly. Inter-observer agreement data should be collected for at least 20% of observations and agreement should be at least 80%.
3. There should be at least three attempts to show an effect of intervention (three phase changes). SCDs meeting these criteria include ABAB design, multiple baseline design with at least three subjects, and alternating treatment design with at least three treatment changes.
4. For multiple baseline designs, there should be a minimum of six phases and three points in each phase to meet WWC standards with reservations and five points in each phase to meet WWC standards (Kratochwill et al., 2010).

Horner et al. (2005) provides some additional guidelines on SCD methodology. The independent variable should be described so that it is replicable and have treatment integrity data. The dependent variable should have social significance. Regarding baseline, it should continue until the data trend is flat or in the direction opposite of what would be expected with intervention (Horner et al., 2005). An intervention is considered evidence-based when the methodology meets the above standards, a causal relationship between the intervention and dependent variable is demonstrated by three replications of effect, and the effect is replicated across studies. Finally, strength of effect should be determined using visual analysis and calculation of effect sizes (Katochwill et al., 2010). SCD is useful because it does not require large sample sizes like group design while still guarding against threats to internal validity such as selection, events occurring at the same time as intervention, maturation of subjects, statistical regression, repeated testing,

and observer drift. However, external validity or generalizability is still limited to the particular subjects and settings in the study (Kratochwill et al., 2010). Another limitation of SCD is effect sizes estimated from SCD are typically larger than those found from group designs. This is due to the autocorrelation of repeated measures, which yields less variability within subjects relative to between subjects and therefore a larger effect size (Maggin et al., 2011).

Setting and Participants

Participants were recruited from four public elementary schools in an urban city in Southern California. The school district has a total of 22 elementary schools with mostly minority students. Based on most recent data available, the school has a Hispanic/Latino population of 78% and an African American population of 16.6%. Sixty-two percent of students are English Language Learners, 5.6% percent qualify for special education services, and 84% are from socio-economically disadvantaged backgounds (e.g., qualify for free or reduced lunch). On recent universal screening reading assessments, about 25% of the students scored at-risk. Less than 50% of students scored in the low-risk category. The remaining students scored in the some-risk category. They district's elementary schools typically stratify classes in each grade into one higher performing class and remaining lower performing classes. In each class, there are students who scored in the at-risk, some-risk, and low-risk ranges, but the higher performing classes have larger percentages of students who scored low-risk.

The participants included four certified, general education teachers in grades 2 to 5 and the students in their classrooms. Participants were selected from four schools that

were implementing PALS. Only teachers in grades 2 to 5 who had not implemented PALS during past school years were considered for inclusion in the study. A volunteer list of eligible classrooms was generated from each school, and then one classroom from each school's list was randomly selected to participate in the study.

The participating classrooms included one grade second grade classroom, one third grade classroom, and two fifth grade classrooms. The second grade classroom was a high performing classroom, with an average oral read fluency score of 122 words per minute prior to the start of the student. The third grade classroom was a lower performing classroom, with an average oral reading fluency score of 59 words per minute prior to the start of the study. The number of students in each classroom ranged from 17 to 32. Table 1 shows the number of students in each class, the average classroom oral reading fluency score prior to PALS, and the percentile rank the score corresponds to on Aimsweb national norms that grade level. In each participating classroom, three ELLs and three native English speakers with the lowest reading performance on screening measures were progress monitored.

The consultants in this study were district-hired RTI Specialists who were doctoral graduate students in school psychology. The consultants were trained in implementing RTI, administering standardized assessments, using direct observation checklists, and consulting with teachers. They were also trained on PALS procedures and have had experience conducting PALS trainings and assisting with implementation.

PALS Reading Intervention

Peer Assisted Learning Strategies (PALS) is a class-wide peer tutoring intervention developed by Lynn and Doug Fuchs. PALS for grades 2 to 6 focuses on reading fluency and comprehension skills. PALS should be implemented at least three days a week and each session requires about 30 minutes. PALS provides a manual with materials and scripted training lessons for the teacher to teach students the PALS procedures and activities. To partner students in PALS, students are ranked by reading performance and split into a higher and lower group. They are then paired so that the first ranked reader from the higher group is paired with the first ranked reader in the lower group. The second ranked reader for the higher group is paired with the second ranked reader in the lower group and so forth. Students in each dyad take turns being the peer tutor (coach) and tutee (reader). The higher reader of the pair is designated the first reader and the lower reader is designated the second reader. In addition to the reading activities, PALS incorporates a points system to provide positive reinforcement to students (Fuchs et al., 2008).

Students engage in the following four activities in order: partner reading, retell, paragraph shrinking, and prediction relay. In the first activity, partner reading, the first reader reads for five minutes while the second reader corrects mistakes using a scripted correction procedure and marks a point for each sentence read. Then, the students switch roles and the second reader reads from where the first reader began, also for five minutes. In retell, for two minutes, the first reader prompts the second reader to retell the events that happened in the text read during partner reading. In paragraph shrinking, the first

reader reads one paragraph at a time and makes a main idea statement after each paragraph. The second reader gives prompts and marks points. After five minutes, the students switch roles. Finally, in prediction relay, the second reader prompts the first reader to make a prediction, read half a page, and decide if the prediction was correct. Again, the second reader marks points and students switch roles after five minutes (Fuchs et al., 2008).

Materials

The PALS manual for grades 2 to 6 consists of student and teacher materials and 12 scripted training lessons that teachers use to teach students the PALS procedures and activities. The 12 training lessons are sequential and each builds upon what was taught in the previous training lessons until all four PALS activities are taught. The training lessons continuously review how to conduct previous activities so there are opportunities for teachers and students to self-correct aspects of implementation previously omitted or conducted incorrectly. After the 12 training lessons are completed, teachers are instructed to use the teacher command card to implement all four activities in subsequent PALS sessions (Fuchs et al., 2008).

Student materials for each pair of students included one set of question cards for each PALS activity, a correction card, and point sheets. PALS manuals and student materials were provided to teachers prior to PALS implementation. PALS books were at the reading level of the lower reader in each pair, and were obtained from the school or classroom library.

Measures

Treatment Integrity Checklist. The Heartland Area Education Agency (2006) adapted a direct observation treatment integrity checklist developed by the PALS authors. The checklist is for full PALS lessons including all four PALS activities. A treatment integrity percentage score is calculated from a detailed list of teacher and student behaviors. As recommended by Noell and colleagues (2000), studies on performance feedback for peer tutoring should measure both teacher behaviors and student behaviors. Furthermore, since teachers are responsible for teaching students the PALS procedures, student treatment integrity is a reflection of teacher treatment integrity.

The checklist has a total of 99 items, which are scored as observed or not observed. The first items on the checklist pertain to classroom arrangement and the presence of necessary teacher and student materials. General teacher behaviors listed include whether the teacher monitored student pairs, gave positive feedback and points for good behavior, gave corrective feedback, and correctly timed each activity. Student behaviors are observed for two student pairs during each of the four activities. Student items are precise and correspond to steps in the training lessons. For example, the first three steps for Partner Reading are "reader 1 reads aloud from book for 5 minutes," "reader 2 corrects mistakes using the correction procedure," and "reader 2 awards 1 point for each correctly read sentence." (Heartland Area Education Agency, 2006, p. 1). The items observed for the teacher and for each pair during each activity are summed and divided by the total number of items observable to obtain an overall treatment integrity percentage score (Heartland Area Education, 2006). The checklist follows the primary

approach to creating checklists described in the literature, which is to describe steps so that they are observable and have enough specificity and sensitivity to show a relationship between levels of integrity and outcomes (Noell, 2010). Based on pilot data, inter-observer reliability for the checklist was 95%. See Figure 2 for the full checklist.

Consultant Procedural Integrity Checklist (CPI). A consultant procedural integrity checklist was created to ensure that all steps of the performance feedback procedure are implemented each time. The checklists instructs that for each performance feedback day, the consultant provide positive and corrective feedback as described in the procedures section, show the teacher the treatment integrity data, then complete a new treatment integrity observation. Every two weeks, the consultant was directed to review student progress monitoring data with the teacher. See Figure 3 for the checklist.

Oral Reading Fluency (ORF). ORF is a standardized, general reading outcome measure of oral reading fluency. It is used for screening three times a year and for monitoring student progress. ORF is individually administrated. The examiner presents a reading passage of a specific grade level to the student and asks the students to read aloud. If the student does not read a word in three seconds, the examiner tells the student the word. No other corrections are made. The student's score is the number of words read correctly in one minute (Shinn & Shinn, 2002a). ORF for grades 2 to 5 has been found to correlate well with end of year measures of general reading achievement, with correlation coefficients of approximately .70 (Pearson, 2012). For grades 2 to 4, ORF is a better predictor of criterion reading assessments compared to MAZE comprehension. For grade

5, ORF is about as good a predictor of criterion reading assessments as MAZE is (Wayman, Wallace, Ticha, & Espin, 2007).

Grade 2 to 5 ORF probes are sensitive to growth and single probes have high alternate-form and split-half reliability coefficients of .94, so single ORF probes have sufficient reliability for screening, progress monitoring, and individual decision making (Pearson, 2012; Salvia & Ysseldyke, 2007). High alternate-form reliability also indicates equivalency of passages, which is desirable when using one probe instead of the median of three probes. ORF probes are as reliable for low performing students as they are for higher performing students. Shinn, Gleeson and Tindal (1989) found standard error of estimates (SEEs) did not differ for passages that are more or less difficult for a student. This indicates that for a lower performing student, for whom grade level passages would be more difficult than for a higher performing student, SEEs are not different than SEEs for a higher performing student. In other words, it indicates that lower performing students do not exhibit more variability on CBM probes compared to higher performing students (Shinn et al., 1989). In addition, Deno and colleagues (2001) found that standard errors of the slopes (SE_b) were smaller for students with lower initial rates of performance. Lastly, frequent and longer durations of progress monitoring make the growth rate more reliable. Christ (2006) found that the SE_b was 9.19 for 2 weeks but only .42 for 15 weeks, and nine to ten data points reduce SE_b to below 1. Inter-observer reliabilities are .99 (Pearson, 2012).

Maze Comprehension (MAZE). The MAZE is a standardized, group administered, measure of reading comprehension. It is used for screening three times a

year and may be used for progress monitoring comprehension. On the MAZE, after the first sentence, every seventh word is replaced with three words in parenthesis. The student is instructed to silently read the story and circle one word in each parenthesis that makes the most sense, for three minutes. The student's score is the number of correct answers (Shinn & Shinn, 2002b). The correlations between MAZE and end of year tests for grades 3 to 5 are .59. Alternate form reliabilities from fall to winter for grades 2 to 5 range from .68 to .78. Since the time between administrations was four months, these alternate form reliabilities estimates are conservative (Pearson, 2012).

Consultant Rating Profile. A consultant rating profile adapted from the consultant rating profile in Noell et al. (2005) was used as a measure of social validity. The term consultant was changed to "RTI Specialist," two of the questions were slightly adjusted to be specific to the present study, one question was changed, and the order of questions was adjusted. The first seven items pertain to the effectiveness of the consultant and performance feedback procedure. The last three items pertain to the effectiveness of the PALS intervention. Teachers are asked to rate their agreement with each item on a seven point Likert scale (1 = Strongly Disagree, 7 = Strong Agree). For the first seven items, teachers are asked to rate their agreement for the period of time the RTI Specialist provided feedback. Internal consistency of Noell et al.'s (2005) consultant rating profile as measured by Cronbach's α was .89. Cronbach's α is similar to split-half reliability and can be understood as the average split-half correlation based on all possible divisions of the items in half (Salvia & Ysseldyke, 2007). In addition, high internal consistency indicates that the items can be considered as one scale (Noell et al., 2005). In Noell et

al.'s (2005) study, consultants were rated 6.5 out of 7 on average. See Figure 4 for the rating scale.

Inter-observer Agreement

Inter-observer data were collected for 20% of the treatment integrity observations and for 20% of the ORF progress monitoring measures. The second observer was another RTI Specialist. For inter-observer observations using the treatment integrity checklist, both observers watched the same pairs of students during each PALS activity. Inter-observer agreement for the treatment integrity checklists was the number of agreements divided by the total number of items observed. Inter-observer agreement for ORF was the number of agreements divided by the total number of words the student read in one minute (Pearson, 2012). Inter-observer agreement was 94%.

Procedure

Initial teacher training. Prior to the start of PALS implementation, an RTI Specialist provided teachers with a 30-minute power-point training presentation. The presentation presented an overview of PALS, including descriptions of the purpose and benefits of PALS, the purpose of the training lessons, how to pair students, how to select books for PALS, the four partner activities, and the student materials. It was stated that PALS should be implemented at least three days a week. Teachers' questions and concerns were addressed.

Implementation of training lessons. Teachers implemented the 12 PALS training lessons using the scripted procedures in the manual. The consultant did not provide feedback during this time. All 12 training lessons need to have been completed

prior to measuring baseline treatment integrity, because the first research question is to what integrity teachers implement all four PALS activities when provided with a training presentation, materials, and the manual. As stated above, the training lessons provide opportunities for teachers and students to self-correct implementation errors from previous lessons, so implementation of PALS after all training lessons have been completed would best represent how well the teachers implement PALS using the resources provided.

Baseline. During the baseline phase, twice a week, the RTI consultants measured treatment integrity of full PALS lessons using the direct observation treatment integrity checklist. Baseline for each classroom continued until the data showed a stable flat or downward trend. Baseline included a minimum of three points. Classrooms for which average treatment integrity during baseline was less than 70% and stable or downward trending participated in the performance feedback phase. In the studies by Witt et al. (1997), Noell et al. (1997), and Mortenson and Witt (1998), performance feedback was provided when average baseline was below 73% and stable or downward trending. A negative slope during baseline indicates treatment integrity will decline over time without intervention, so provision of performance feedback is warranted to increase integrity (Solomon et al., 2012). In Solomon and colleagues' (2012) meta-analysis, the average trend of baseline was -2.83, indicating that treatment integrity does tend to decline without support. All four classrooms met the above criteria to continue into the performance feedback phase.

As characteristic of multiple baseline design, introduction of performance feedback was staggered. After the first class entered the performance feedback phase, the second class continued in baseline for at least three more observations before receiving performance feedback. After the second class entered the performance feedback phase, the third class continued in baseline for at least three more observations and then received performance feedback. After the third class entered the performance feedback phase, the fourth class continued in baseline for at least three more observations and then received performance feedback (Horner et al., 2005).

Performance feedback. During the performance feedback phase for each classroom, a consultant continued to observe full PALS lessons and measure integrity two times a week. The percentage of implementation integrity achieved at each observation was recorded on a graph with previous integrity scores. Prior to the start of that day's PALS session, the consultant met with the teacher for 3 to 5 minutes to provide feedback on the previous PALS session observed (Noell et al., 1997, Mortenson & Witt, 1998; Witt et al., 1997). In the performance feedback procedure, the consultant named strengths or improvements and provided corrective feedback on missed items (behaviors). Items corresponding to key components of PALS as identified by Vadasy and colleagues (1997) took precedence. See Table 3 for the list of key components. The teacher was asked to remind the entire class during the current day's PALS lesson about missed items discussed. In addition, the consultant showed the teacher a graph of the treatment integrity data collected thus far.

For use in performance feedback to teachers, student progress on ORF was collected weekly and shared with the teacher every two weeks. Progress monitoring data aids teachers in making instructional decisions for students (Fuchs & Fuchs, 1986). For promoting treatment integrity, providing teachers with an objective measure of student growth may encourage them to continue implementing the intervention with fidelity when students make process (Gersten et al., 2000) or increase their perceptions of student need when progress is inadequate (Shinn, 2005). Teacher perceptions of intervention effectiveness and student need are both factors affecting implementation (Durlak & Dupre, 2008; Gresham, 1989).

Provision of CBM progress monitoring data to teachers is most necessary for low performing students, as teacher judgment without CBM data is least accurate for low performing students (Begeny, Krouse, Brown, & Mann, 2011; Madelaine & Wheldall, 2005). In the study by Begeny and colleagues (2011), teachers over-estimated and under-estimated students' risk levels and on average over-estimated words read per minute. The extent of inaccuracy was greater for at risk and some risk students compared to for low risk students. In a study by Graney (2008), teacher judgments of growth rate over a six-week period for low performing students were not correlated with actual growth rates measured by CBM. In the present study, progress monitoring data was provided for low performing students for the above reasons, and because PALS has been found to be equally effective for low and average performing students (Fuchs et al. 1997). Since progress monitoring data was collected for three lowest ELL and three lowest native English speaker students in each class for the additional purpose of examining student

outcomes, data for all six students was shared with the teacher. When used as part of performance feedback in actual school practice, Shinn (2005) suggests that providing progress data for one low student is sufficient, as his study found that providing progress data for an individual or a group of low performing students did not differently affect student growth rates.

Finally, the consultant completed the consultation procedural integrity checklist during each performance feedback session. The section on the checklist pertaining to progress monitoring was completed once a week. Permanent products collected included treatment integrity checklists, progress monitoring data, and graphs of treatment integrity scores.

Progress monitoring. One purpose of progress monitoring was to provide feedback on student progress as part of the performance feedback procedure. A second purpose was to gather student data to examine the impact of PALS and performance feedback on student growth rates. In each class, three ELLs and three native English speakers who scored the lowest on most recent reading screening measures were progress monitored weekly during the baseline and performance feedback phases. A composite of AIMSweb ORF and MAZE comprehension scores was used to identify the lowest students.

Students were progress monitored with AIMSweb ORF probes. The grade level of probes used was the highest level up to grade level on which the student could read at least 25 words per minute. For example, if a student was in 4^{th} grade, the examiner started with 4^{th} grade passages. If the student could not read at least 25 words correctly in one

minute, the examiner then tested back to 3rd grade passages and so forth until the student could read at least 25 words per minute.

One probe at the identified level was administered each week. Each student's scores were recorded in the AIMSweb database's progress monitoring schedules. The progress trend line was graphed with a goal line, which is a line connecting the student's baseline score to the goal score. Goals were set using end of year grade level benchmark goals. These are the cutoff scores for the low risk category at the end of the year benchmark period. They are 92 words per minute for grade 2, 119 words per minute for grade 3, 136 words for minute for grade 4, and 143 words per minute for grade 5 (Pearson, 2011).

Analyses

Research question 1. For each classroom, the mean treatment integrity achieved during the baseline phases was compared to a cutoff of 80% for high treatment integrity. An 80% cutoff was chosen for the following reasons. In Hagermoser and Kratochwill's (2008) study, most published studies that measured treatment integrity reported at least 80% integrity. Second, in the studies by Fuchs and colleagues (1997) and Saenz and colleagues (2005) conducted with PALS for grades 2 to 6, the average integrity for teacher and student behaviors was 80% to 95%. Based on previous research (e.g., Witt et al., 1997; Vadasy et al., 1997), it was hypothesized that during baseline, average treatment integrity will be below 80%.

Research question 2: The standards for conducting SCD studies compiled by Kratochwill et al. (2010) for WWC state that analysis of SCD results should begin with

visual analysis, followed by estimation of effect size. In this study, the impact of performance feedback on treatment integrity was evaluated using visual analysis and four effect sizes. The effect sizes calculated were percent of all non-overlapping data (PAND), improvement rate difference (IRD), R^2, and Pearson's *Phi* coefficient.

Visual analysis. The visual analysis process began with determining that the sets of baseline data have a predictable trend. Next, within-phase trend and variability of data during the performance feedback phases were examined to determine the consistency of intervention effects. Third, each performance feedback phase was compared to each baseline phase to determine if performance feedback produced an effect. At this step, the immediacy of changes in level, trend, and variability from the last three points of baseline to the first three points of the performance feedback phase was examined. In addition, the amount of overlap between performance feedback and baseline phases was examined. Lastly, the above information was synthesized to determine if there were three replications of effect to warrant concluding that there is a functional relationship between the independent variable (performance feedback) and the dependent variable (treatment integrity). At this step, consistency of data patterns across classrooms was examined (Kratochwill et al., 2010).

Effect sizes. Desirable features of an effect size for single case design include consistency with the logic of visual analysis, ease of interpretation by researchers across fields, estimation of the magnitude of effect, ability to control for autocorrelation and within phase trends, and distributional characteristics based on statistical theory. There is yet to be agreement on which method of effect size estimation for SCD is best because

each has advantages and disadvantages (Maggin et al., 2011). Since researchers use various effect size estimation methods and not one method fulfills all criteria desirable for a SCD effect size, several leading methods were used in data analysis for research question 2.

PAND. Non-parametric methods such as Percent of Non-overlapping Data (PND) and Percent of All Non-overlapping Data (PAND) are commonly used and simple ways of estimating effect size in single case design (Kratochwill et al., 2010). They quantify the amount of data in the intervention phase that does not overlap with data in the baseline phase. A main difference between PND and PAND is PAND includes all data points and thus takes into account length of baseline (Parker, Hagan-Burke, & Vannest, 2007). An advantage of non-parametric methods is they do not require the assumption of independence of observations, which is not met in single case design. However, they have a number of weaknesses. First, the effect size is sensitive to outliers in the baseline phase because intervention points are compared to the highest or lowest baseline point. Second, different magnitudes of effect can have the same PND and PAND. Third, because they do not depend on a normal distribution, distribution of the sample is unknown. Finally, since non-parametric effect sizes do not account for autocorrelation or non-independence of data, they can result in inflated effect sizes, which then cannot be directly compared to group design effect sizes (cohen's *d*). Of the non-parametric effect sizes, PAND is the most supported because it uses all data points and can be converted to a Pearson *Phi* (Φ) coefficient, which has a known sampling distribution from which p values and confidence intervals can be calculated (Maggin et al., 2011).

To calculate Percent of All Non-overlapping Data (PAND), the number of overlapping points for each classroom were counted and summed. Parker et al. (2007) defines overlapping data points as "the minimum number of points that would have to be swapped across phases for complete score separation" (p. 197). PAND = 100% - total number of overlapping points / total number of observation points (Parker et al., 2007). In the study by Parker and colleagues (2007) of 75 multiple baseline designs, interventions that were 50th percentile in terms of effectiveness had a PAND of about 84%. In other words, interventions that yielded a PAND of 84% were more effective than half of the interventions compared to. PAND for 25th percentile and 75th percentile interventions was 72% and 92% respectively. Interventions with 50% PAND or less are considered ineffective, because chance overlap would produce 50% PAND. These values were used to guide interpretation of PAND for the effectiveness of performance feedback on treatment integrity (Parker et al., 2007).

IRD. Improvement rate difference (IRD) is another useful, non-parametric effect size for single case research that is calculated from non-overlapping data (Maggin et al., 2011). It is the difference between the improvement rates of baseline and intervention phases (Parker, Vannest, & Brown, 2009). IRD was used in the meta-analysis of performance feedback studies by Solomon et al. (2012) and has been long used as a "risk difference" in medical studies (Parker et al., 2009, p. 138). IRD has a known sampling distribution so confidence intervals can be calculated (Parker et al., 2009). Parker and colleagues (2009) found that the correlation between IRD and R^2 was .86, though at the

higher end of the distribution of interventions ranked by effectiveness, IRD values were higher than R^2 and had a ceiling effect.

To calculate IRD, the improvement rate during baseline was subtracted from the improvement rate during performance feedback. The improvement rate in baseline is the number of baseline points higher than any point in the performance feedback phase, divided by the total number of baseline points. The improvement rate in performance feedback is the number of performance feedback points higher than all baseline points, divided by the total number for performance feedback points. IRDs for individual classrooms were calculated first, then averaged to obtain an overall IRD. For individual classrooms, 95% confidence intervals were calculated using a test of two proportions (Parker et al., 2009). Based on 166 published sets of AB data, an IRD of .48 was at the 25^{th} percentile, an IRD of .71 was at the 50^{th} percentile, and an IRD of .90 was at the 75^{th} percentile.

R^2. Kratochwill and colleagues (2013) recommended including regression-based effect sizes due to their technical qualities and practicality. R^2 is a parametric effect size estimate, meaning it is regression-based and assumes the data have a normal distribution. It uses all data points, and accounts for level, trend, and variability of data. Since the R^2 coefficient has a known sampling distribution, *p* values for hypothesis testing and confidence intervals can be calculated (Parker et al., 2007). R^2 can be interpreted in SCD as the proportion of score variance that is explained by phase differences (Cohen, 1988). The main limitation of R^2 is the need for regression assumptions of homogeneity of variance, normality, and independence of data (Parker et al., 2007). In SCD design,

independence of data is particularly not met due to autocorrelation and autocorrelation is difficult to model (Maggin et al., 2011). Another limitation is a large sample size is needed to best calculate parametric effect sizes (Maggin et al., 2011).

Analysis of variance was used to calculate the R^2 effect size of performance feedback on treatment integrity. Treatment integrity scores in all baseline observations was compared to treatment integrity scores in all performance feedback phase observations (Parker & Hagan-Burke, 2007). For 25th percentile, 50th percentile, and 75th percentile interventions in the Parker et al. (2007) study, R^2 was 0.22, 0.50, and 0.67 respectively. In an analysis of a single case data set by Parker and Hagan-Burke (2007), R^2 was .17 when Cohen's percent of non-overlapping data was 50%, which would indicate an ineffective intervention. R^2 was .50 when data non-overlap was 80%. R^2 was approximately .64 when data non-overlap was 90%, which corresponds to an effective intervention. The R^2 values for ineffective, moderately effective, and effective interventions were similar in Parker & Hagan-Burke (2007) and Parker et al., (2007). Thus, for single case design, an R^2 of approximately .2 in single case design may be considered small, an R^2 of .50 may be considered medium, and an R^2 of approximately .65 may be considered large.

Φ. A leading bona fide, alternative effect size to R^2 that can be calculated from PAND is Pearson's *Phi* (*Φ*). Like R^2, *Φ* has a known sampling distribution so *p* values and confidence intervals can be calculated. Φ^2 and R^2 are correlated at 0.90, indicating that they measure similar constructs. An advantage of *Φ* over R^2 is *Φ* does not require assumptions of homogeneity of variance, normality, or independence of data. One

disadvantage is Φ has less statistical power compared to R^2. In a study comparing Φ^2 and R^2 using 75 multiple baseline designs, Φ^2 detected effects as low as 0.34 while R^2 detected effects as low as 0.10. Φ also produced larger effect sizes than R^2 at the higher end of the distribution of effect sizes and lower effects than R^2 at the lower end of the distribution (Parker et al., 2007).

To calculate Φ, a 2 x 2 table was used. See Table 2. Φ was calculated for all classrooms in the multiple baseline design considered together. First, the percentage of baseline and performance feedback points was entered in the marginal total columns. Second, the percentage of performance feedback points overlapping with the highest baseline points was calculated and divided between the cells *b* and *c*, which represent performance feedback points lower than baseline and baseline points higher than performance feedback. Next, cell *a* was calculated by subtracting cell *c* from the percentage of baseline points. Cell *d* was calculated by subtracting cell *b* from the percentage of intervention points. Finally, Φ was calculated using the following formula: $[a/(a+c)] - [b/(b+d)]$ (Parker et al., 2007). For 25th, 50th, and 75th percentile interventions in the Parker et al. (2007) study, Φ^2 was .22, .53, and .80 respectively.

Research question 3. For 17 main components of PALS, treatment integrity during baseline observations and performance feedback observations was examined and compared. Key components identified by Vadasy and colleagues (1997) included: pairing of students, higher reader reading first, completion of the partner reading, paragraph shrinking, and prediction relay activities, correct timing of each activity, use of question cards, use of the points system, teacher monitoring of pairs, and provision of corrective

and positive feedback to students. Additional main components in PALS are use of the correction procedure and completion and correct timing of the retell activity. See Table 3 for the list of components. The percentage of baseline observations in which the component was observed and the percentage of performance feedback observations in which the component was observed was calculated. It was hypothesized that treatment integrity of key components would be greater during performance feedback than during baseline.

Research question 4. Progress monitoring data were used to examine the effect of performance feedback on student growth rates for ELLs and native English Speakers. Since ORF is a general outcome measure, intervention may not yield an immediate jump in level of performance, but effective intervention should increase growth rates (Pearson 2012, Vaughn et al., 2010). Other multiple baseline tutoring studies that have used ORF as an outcome measure have focused on growth rates (e.g., Fiala & Sheridan, 2003; Persampieri, Gortmaker, Daly III, Sheridan, & McCurdy, 2006).

First, progress monitoring data for ELLs and native English speakers were displayed using a multiple baseline graph, with average ORF score on the vertical axis and one line representing each group. For each classroom, ORF scores at each progress monitoring time were averaged. Next, the visual analysis procedure described above was used to examine the impact of performance feedback on ORF. Of main interest on this multiple baseline graph was whether trend (rate of growth) increased from baseline to performance feedback. PAND was calculated separately for ELLs and native English speakers.

Additionally, matched pairs, one-sample t-tests were used to examine if growth rates during performance feedback were significantly higher than growth rates during baseline. Whereas visual analysis allows for qualitative inspection of changes in growth rates, t-tests allow for significance testing. Three t-tests were conducted: one for ELLs only, one for native English speakers only, and one for ELLs and native English speakers combined. For each t-test, each student's growth rates during baseline and performance feedback phases were calculated using least squares regression lines. Next, the differences in growth rates between phases were calculated. The null hypothesis was the average difference (μ) is zero. The alternate hypothesis was the average difference is greater than or less than 0. The t-statistic was calculated as shown below.

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

$$t = \frac{\bar{\chi} - 0}{s/\sqrt{n}}$$

Degrees of freedom was 23 (n-1) for ELLs and native English speakers combined, 11 for ELLs only, and 11 for native English speakers only (Raykov & Marcoulides, 2013).

A power analysis indicated that for a two-tailed, matched-pairs t-test, a sample size of 24 with α level .05 can detect with .80 power an effect size of .59, which is a medium effect. A sample size of 12 with α level .05 can detect with .80 power an effect size of .89, which is a large effect (Cohen, 1988). In Noell et al.'s (2005) group study, the effect size of performance feedback on student outcomes was $\eta = .36$, which is a large effect.

Research Question 5. The consultant rating profile was used to examine teachers' acceptability of the consultant and performance feedback procedure, and their perception of PALS' effectiveness. Teachers' rating forms were examined individually first, since performance feedback and PALS may be more acceptable to some teachers than others. Then, the average rating for each item was calculated. It was hypothesized that teachers will be satisfied with the consultant, performance feedback procedure, and PALS.

Results

Research question 1

During baseline, all classrooms achieved less than 80% average treatment integrity. Average treatment integrity for classrooms 1, 2, 3, and 4 was 13%, 70%, 22%, and 61% respectively. Average baseline treatment integrity across classrooms was 42%.

Research question 2

Visual analysis. During baseline, treatment integrity was downward trending or flat for each class. During performance feedback, treatment integrity was steadily upward trending for each class, and the increase in level and slope was immediate. Table 4 displays the mean treatment integrity achieved during performance feedback compared to the mean treatment integrity achieved during baseline for each classroom. It also displays each classroom's final treatment integrity. There was no overlap of percent treatment integrity data points between phases for classrooms 1 and 3, and very little overlap of data points between phases for classrooms 2 and 4. Variability was similar during the last three points in baseline and the first three points of performance feedback. The data

patterns were consistent across classrooms and indicate four replications of effect and therefore a functional relationship between performance feedback and treatment integrity. See Figure 5 for the multiple baseline results.

Effect sizes. PAND calculated across classrooms was 93%, which indicates an effective intervention. In Parker et al. (2007), an intervention with 93% PAND was at the 75th percentile or higher in terms of effectiveness. The corresponding Φ was .88 and Φ^2 was .77. A Φ^2 of .77 indicates an effective intervention, as a Φ^2 of .80 was at the 75% percentile in terms of effectiveness in Parker et al. (2007).

IRDs and corresponding 95% confidence intervals for each classroom are displayed in Table 5. The IRD averaged across classrooms was .73, which in the Parker and colleagues (2009) study was at approximately the 50th percentile of effectiveness. IRD for classrooms 1 and 3 however, were 1.00, indicating the intervention was very effective for those classrooms. The 95% confidence intervals for the IRDs indicate that the true effect was small to large.

The R^2 effect sizes and corresponding 95% confidence intervals for each classroom are also displayed in Table 5. The average R^2 was .51, which was at the 50th percentile for effectiveness in the studies by Parker and colleague (2007) and Parker and Hagan Burke (2007). The 95% confidence intervals indicate the true R^2 effect sizes were small to large (Parker et al., 2007; Parker & Hagan Burke, 2007).

Research question 3

Table 3 displays the percentage of baseline and performance feedback observations in which each core component was implemented. Of the 17 core components, eight were implemented during at least 80% of baseline observations. They were as follows: students were paired, students were seated side by side, the higher reader read first, partner reading was done, retell was done, prediction was done, teacher monitored students, and teacher provided corrective feedback. Of the 17 core components, 14 were implemented in at least 80% of performance feedback observations. The three components that were still implemented less than 80% of the time were use of the correction procedure, use of the points system, and correct timing of prediction relay.

Research question 4

Visual Analysis. Figure 6 displays the multiple baseline graph of average ORF during baseline and performance feedback. Overall, ORF scores for native English speakers were higher than scores for ELLs, though classroom 3 shows overlap between the two groups. For both ELLs and native English speakers in classrooms 1, 2, and 4, ORF during baseline was flat or downward trending. For both groups in classroom 3, ORF was slightly upward trending during baseline.

For native English Speakers in all four classrooms, ORF was upward trending during performance feedback. The increase in trend from baseline to performance feedback was immediate to delayed by one week. There was an increase in level in classrooms 3 and 4, though there were few points in the performance feedback phases. In classrooms 1 and 2, variability of data decreased from baseline to performance feedback.

PAND was 81%, indicating about 20% data overlap. Overall, the data patterns indicate at least three replications of effect and suggest a functional relationship between performance feedback and reading growth for native English Speakers. In Parker et al. (2007), a PAND of 81% would have been at about the 35th percentile for effectiveness.

For ELLs, ORF was upward trending during performance feedback in classrooms 1, 2, and 4. Although the positive slope in classroom 2 was slight, slope during baseline had been negative. The increases in trend were immediate to delayed by one week. Although there were changes in trend, there was little increase in level (average ORF) from baseline to performance feedback phases. Variability of data was similar across phases. There was more data overlap between phases for ELLs compared to for native English speakers (64% PAND). The data patterns in classrooms 1, 2, and 4 indicate that there *may* be three replications of effect and a functional relationship between performance feedback on reading growth for ELLs. However, the effect for ELLs is small, as a PAND of 64% is at approximately the 15th percentile for effectiveness (Parker et al., 2007). Furthermore, the data was interpreted with caution due to the limited number of progress monitoring points in some phases.

T-tests. Table 6 displays the average growth rates and corresponding standard deviations during baseline and performance feedback phases for students who were progress monitored. Growth rates are displayed for ELLs, native English speakers, and ELLs and native English speakers combined. The mean growth rate during baseline was 1.66 words per week for ELLs and 2.47 words per week for native English speakers. The mean growth rate during performance feedback was 2.29 for ELLs and 3.47 for native

English speakers. Table 6 also displays the results of each t-test conducted. For each group, average growth rates during the performance feedback phase were higher than those during baseline, but the differences were not statistically significant ($p > 05$).

Research question 5. Three of the four teachers were satisfied with the performance feedback procedure, treatment integrity graphs, progress monitoring graphs, and PALS program. One teacher did not agree that performance feedback was a good use of time and did not perceive PALS as effective. Table 7 displays the average rating across teachers for each item on the consultant rating profile. Average ratings ranged from 5 to 7 (1 = strongly disagree, 7 = strongly agree). The average ratings for the last two items pertaining to PALS' effectiveness were slightly lower than the items pertaining to performance feedback.

Discussion

The primary purpose of the study was to examine if there is a functional relationship between performance feedback and overall treatment integrity when performance feedback is used for a standard protocol, class-wide, reading intervention. Additional purposes were to examine the impact of performance feedback on the implementation of core intervention components and student growth rates, and to examine teacher acceptability of the performance feedback procedure. The following sections will address the significance of the study's findings and possible explanations for the results. First, the impact of performance feedback on treatment integrity and student growth rates will be discussed. The impact of performance feedback on overall treatment integrity will be evaluated based on the visual analysis data and effect sizes. The impact

of performance feedback on the implementation of core intervention components will be discussed and compared to the study conducted by Vadasy and colleagues (1997). Regarding the impact of treatment integrity on student growth, possible reasons for the lack of significant differences between baseline and performance feedback growth rates will be explored. Then, teachers' acceptability of the performance feedback procedure, the consultants, and PALS will be discussed. Finally, limitations of the study, implications for practice, and directions for future research will be considered.

Treatment integrity without performance feedback

Prior to the provision of performance feedback, classrooms achieved less than the desired 80% treatment integrity (Durlak & Dupre, 2008). Baseline treatment integrity ranged between 13% and 70%. Two classrooms implemented PALS with low integrity and omitted or incorrectly implemented many of the core components, including question cards, correction cards, point sheets, correct timing of each activity, correct amount of text read, and teacher monitoring of student pairs. As studies have found, interventions implemented with higher integrity produce better outcomes and low treatment integrity can result in minimal improvement or negative outcomes (Durlak & Dupre, 2008; Noell, 2010). The other two classrooms implemented PALS with substantially higher integrity, though their average integrity did not reach 80%. Regarding the PALS intervention, a possible hypothesis is the many teacher and student behaviors required by PALS can be difficult to implement without support. As Gresham (1989) stated, interventions that are complex and require more time are less likely to be implemented with integrity.

From examining average student reading levels in each classroom, there does not appear to be a relationship between initial reading levels and baseline treatment integrity. Average ORF for two classes with higher baseline integrity was at the 10th and 45th percentiles. Average ORF for two classes with low baseline integrity was at 85th and 25th percentiles. However, it is notable that the class with highest treatment integrity during baseline and performance feedback phases had the lowest student reading levels (average ORF score at the 10^{th} percentile), but also the fewest number of students. The class with the most students (32) had the lowest baseline and PF TI. Thus, while other factors likely contributed to classroom differences, it seems that it may be easier to monitor treatment integrity when there are fewer students in the class.

Impact of performance feedback on overall treatment integrity

The visual analysis procedure of the multiple baseline data showed a clear functional relationship between performance feedback and treatment integrity. The improvement in treatment integrity when performance feedback was provided was evident in both level and slope, and the effect was replicated in all four classrooms. Table 4 shows that average treatment integrity increased from 42% during baseline to 67% during performance feedback, and reached 80% at the final performance feedback session.

PAND and Φ^2 effect sizes indicated that performance feedback was more effective than about 75% of interventions evaluated using multiple baseline design. In Parker and colleagues' (2007) sample of interventions evaluated using single case design, interventions at the 75^{th} percentile for effectiveness (meaning they were more effective

75% of interventions compared to) produced PAND and Φ^2 effect sizes same as those obtained in this study. IRD and R^2 effect sizes indicated that performance feedback was at approximately the 50th percentile of effectiveness compared to other interventions evaluated using multiple baseline design (Parker et al., 2009; Parker et al., 2007). In Parker and colleagues' (2009) sample of interventions, interventions at the 50th percentile for effectiveness produced IRDs similar to those in this study. Considering these comparisons, performance feedback as used in this study would have been at the 50th to 75th percentile of effectiveness (more effective than 50% to 75% of interventions evaluated using single case design). In addition, the average R^2 of .51 can be interpreted as 51% of the variance in treatment integrity can be attributed to phase differences.

The confidence intervals for the IRD and R^2 effect sizes in this study provided a more conservative and broader estimate; they estimated the effect sizes to be anywhere from small to large. The large confidence intervals can be attributed to the sample sizes from which they were calculated, that is the number of treatment integrity direct observations in each phase for each classroom. The number of data points per phase in single case design is typically small compared to sample sizes in group designs, so they produce larger confidence intervals than those in group designs. In Parker et al.'s (2007) sample of 75 published multiple baseline studies, the average number of data points per phase was 11. In this study, the number of data points per phase ranged from three to 11.

Considering the visual analysis and effect sizes together, there is agreement that performance feedback was effective in increasing treatment integrity. This is consistent with findings from previous performance feedback studies (e.g., Mortenson & Witt,

1998; Noell et al., 1997; Witt et al., 1997). In fact, this study provides stronger evidence in support of performance feedback than previous studies did, because more rigorous multiple case design standards were used, visual analysis showed four replications of effect, and multiple effect size estimates were calculated. As discussed in the literature review, previous performance feedback studies were conducted prior to the establishment of single case design standards and did not estimate effect sizes.

Analysis of individual classrooms showed that the two classrooms (classrooms 2 and 4) that achieved higher treatment integrity during baseline achieved above 80% average treatment integrity during performance feedback. The two classrooms (classrooms 1 and 3) that achieved lower integrity during baseline did not reach 80% average integrity during performance feedback, but they improved substantially and approached 80% treatment integrity. In fact, the net increase in treatment integrity for classrooms 1 and 3 was higher than that of classrooms 2 and 4. Classroom 1 improved slowly at first, and plateaued around 50%, and then increased to above 70%. Classroom 3 improved rapidly and steadily, from 22% average treatment integrity during baseline to 73% at the final session. This suggests that although classrooms with higher initial treatment integrity may achieve higher final integrity, classrooms with low initial integrity can make substantial improvements with performance feedback. Providing performance feedback promptly to classrooms that have low initial treatment integrity can be an efficient and effective use of consultation resources.

A number of factors discussed in the literature and also expressed by the teachers participating in the study may have contributed to the outcomes for individual

classrooms. For classroom 1, the teacher's perception of need (Durlak & Dupre, 2008) may have limited treatment integrity initially. The teacher expressed that the students in her class were high performing and PALS was not flexible or challenging enough for them. She conveyed that the scripted retell, paragraph shrinking, prediction relay questions were too limiting for her students. Second, the perceived effectiveness of PALS as conveyed through student growth data or teacher observation (Gresham, 1989) may have affected treatment integrity. The teacher for classroom 4 expressed that her students were not improving and thus PALS may not be the best intervention for her class. The students who were progress monitored were also the lowest performing in the class, and growth rates on reading fluency were low compared to grade level norms. Third, the many step-by-step prescriptive procedures in PALS and the 30-minute time requirement three days a week may have been too demanding or difficult to implement for some of the teachers (Gresham, 1989). As the treatment integrity direct observation checklists show, classrooms 1 and 3 did not initially use the question cards, correction cards, or point sheets with integrity, and omitted or changed questions. They also shortened the duration of partner reading, paragraph shrinking, and prediction relay in order to include all activities in a 30 minute period or to decrease the duration of PALS altogether. However, these problems were more prevalent during baseline and the first few performance feedback sessions; during the latter performance feedback sessions, as reflected by the treatment integrity data, most components were implemented with integrity. In addition, teacher efficacy and skills may have played a role in individual classrooms. Though not measured in this study, teachers who believe that they can

successfully implement a program and affect student achievement are more likely to implement a program successfully (Ransford et al., 2009). Finally, as reflected by the consultant rating scales, teachers perception and acceptance of the support varied and likely affected implementation outcomes (Noell, 2010). While three of the teachers perceived performance feedback positively, one teacher did not think the performance feedback procedure was a good use of her time and did not perceive the treatment integrity graph as useful. That classroom achieved the lowest average and final treatment integrity.

Impact of performance feedback on core components

During baseline, eight of the 17 core components were implemented in at least 80% of observations, which is similar to what was found in the PALS study by Vadasy and colleagues (1997). In fact, many of the components implemented in less than 80% of observations in the present study were the same components implemented with poor integrity in the study by Vadasy and colleagues (1997), including use of question cards, use of the point system, and correct timing of each activity. This indicates that these components are the least likely to be implemented without performance feedback. Without these core components, it is unlikely that PALS would produce the positive effects seen in the studies by Fuchs and colleagues (1997) and Saenz and colleagues (2005), since treatment integrity in those studies was above 80%.

When performance feedback was provided, the number of core PALS components implemented during at least 80% of observations increased from eight to 14. This indicates that performance feedback targeted at core components successfully

improved implementation of core components. The three components that were still implemented less than 80% of the time were use of the correction procedure, use of the points system, and correct timing of prediction relay. This suggests that these three PALS components may be the most difficult to implement with integrity. The correction procedure required student coaches to use a script to identify and correct their partners' reading mistakes. As observed in this study, often times students did not catch mistakes or gave the word without using the correction procedure. The points system required student coaches to mark points for various steps in the activities. Classrooms 1, 3, and 4 did not implement the points system with fidelity until the final few performance feedback sessions. Finally, classroom 1 often shortened the prediction relay activity due to lack of time. Thus, prediction relay may be implemented with less fidelity because it is the last activity in the session.

Impact of performance feedback on growth rates

Visual analysis of the ORF progress monitoring data suggests moderate evidence for a functional, though small effect of performance feedback on students' reading growth. Compared to a PAND of 93% for the impact of performance feedback on treatment integrity, for ORF, the PAND was 81% and 64% for native English speakers and ELLs respectively. The effect was mainly seen through an increase in trend rather than in level. Increase in trend is important however, as the purpose of reading intervention is to improve student growth rates. The increase in trend for both groups in classroom 1 is particularly notable, and parallels classroom 1's increase in treatment integrity from baseline to performance feedback. In classroom 3, which also improved

substantially in treatment integrity, the trend for native speakers appeared exponential during the performance feedback phase. The positive slopes during performance feedback suggest that with continued use of PALS with high treatment integrity, students reading performance would continue to improve. The two multiple baseline graphs (treatment integrity and ORF) together support the link between performance feedback, treatment integrity, and student outcomes.

The t-tests showed that average growth rates during performance feedback were higher than those during baseline, but the differences were also not statistically significant. Thus, although performance feedback has been shown to significantly increase treatment integrity and student outcomes in previous studies, this was not found in this present study. One possible reason is the outcome measure used in this study was oral reading fluency, which is a global outcome measure. Global outcome measures assess not just the material students were exposed to during the intervention, but also overall reading skills and thus may take longer to show growth than the measures used in previous performance feedback studies (Wayman et al., 2007). In previous studies, the outcome measure was limited to what was taught and practiced that week (e.g. spelling words, targeted assignments) (Greenwood et al., 2001; Witt et al, 1997). Another possible explanation is students were participating in PALS in both baseline and performance feedback phases. Although classrooms received a higher dose of PALS during performance feedback due to improved treatment integrity, they did receive the PALS intervention during baseline. Furthermore, two of the four classrooms achieved close to high integrity during baseline.

The more likely explanation for the lack of significant results is limited power due to small sample size. The power analysis indicated that the samples could detect medium and large effect sizes, but not small effect sizes. In addition, the large within group variation made it more difficult to detect significant between group differences. The standard deviations for baseline and performance feedback phases were 4.57 and 3.79 respectively. The large standard deviations for individual phases were largely due to the smaller number of data points in each phase, which ranged from three to 11.

Another observation was that average growth rates during baseline and performance feedback phases were higher for native English speakers than for ELLs. This finding is in accordance with the National Center for Education Statistics' (2011) report that ELLs tend to perform lower on reading measures than native English speakers. Furthermore, average growth rates for students who were progress monitored in this study were lower than normative growth rates for general education students nation-wide, who are participating in typical instruction (Fuchs et al., 1993). Realistic growth rates as determined by Fuchs et al. (1993) are: 1.5 words per week for grade 2, 1.0 words per week for grade 3, .9 words per week for grade 4, and .5 words per week for grade 5. Across grades, the average was .98 words per week. The overall average growth rate for all students who were progress monitored (ELLs and native English speakers combined) in this study was .65 words per week. This finding is not surprising since the students who were progress monitored were the lowest performing of their classes. Also, 75% of students in the district in this study were eligible for free or reduced lunch, and the National Center for Education Statistics (2011) showed that students of low socio-

economic status tend to perform lower than average SES students. Thus, the lower growth rates found in this study can be explained by the demographic characteristics of the sample.

Teacher acceptability

Overall, teachers were satisfied with the consultants and the performance feedback procedure. The positive ratings may have been in part to due the consultants' preexisting working relationship with the schools and the teachers. Teachers may have viewed the consultants as knowledgeable and regarded them positively. As French and Raven (1959) theorized, expert power and referent power contribute to an individual's ability to influence others. Cialdini (2006) also described liking as a principle of influence.

In fact, the use of district-hired consultants rather than researchers makes this study unique from the majority of previous performance feedback studies (e.g. Noell et al., 1997; Noell et al., 2005; Witt et al., 1997). Only one other multiple baseline study on performance feedback has used internal consultants (Hagermoser Sanetti, Fallon, & Collier-Meek, 2013). In Hagermoser Sanetti et al. (2013), the consultants were a school social worker and a special education teacher who had worked in the school for at least three years. They were trained to monitor treatment integrity and deliver performance feedback for tier two behavior interventions. The teachers perceived the intervention and performance feedback positively, and the consultants were able to manage their responsibilities. The consultants did express though that giving feedback to a coworker made them feel uncomfortable. A main difference between Hagermoser Sanetti et al.'s

(2013) study and the present study is in the former study, performance feedback was delivered when treatment integrity was low rather than regularly, which may explain why increases in treatment integrity were not maintained and levels of treatment integrity were variable.

The two items pertaining to PALS' effectiveness received slightly lower average ratings than the items pertaining to performance feedback. This suggests that there may be barriers to PALS' perceived effectiveness. One teacher that had a high performing class expressed that PALS was too rigid and not challenging enough. Another teacher expressed that the students' oral reading fluency scores were not improving. Thus, it seems these two teachers did not perceive PALS to meet the needs of their students, which is a factor contributing to treatment integrity (Durlak & Dupre, Gresham, 1989). As Gresham (1989) discussed, interventions that produce student behavior change more quickly are more likely to be perceived as effective and implemented with integrity. In fact, the teacher that rated the performance feedback procedure and PALS lowest corresponded to the classroom that achieved the lowest average integrity during performance feedback.

Limitations

The first limitation of the study is the performance feedback procedure was not monitored using direct observation. Instead, consultants used self-report checklists to confirm each piece of the performance feedback procedure (positive and corrective feedback, sharing of the treatment integrity graph, sharing of the progress monitoring graph) was delivered. However, permanent products (PALS direct observation treatment

integrity checklists, treatment integrity graphs, and progress monitoring graphs) were produced and collected, ensuring that those aspects of the procedure were conducted.

A second limitation is the performance feedback phase was not followed by a maintenance phase. During a maintenance phase, treatment integrity would be measured again without provision of performance feedback to determine if improvements in treatment integrity are maintained. Adding a maintenance phase would have given insight to the sustainability of the impact of performance feedback on treatment integrity. According to Kearns and colleagues (2010), level or intensity of support, duration of support, quality of training, and flexibility impact the likelihood that an intervention will be sustained. In this study, an initial training was provided, intensity of support was twice per week, and duration of performance feedback was three to eight weeks. The support provided was intermediate between that provided in the studies by Noell et al. (1997) and Mortenson and Witt (1998). In those studies, treatment integrity during the maintenance phases was close to performance feedback levels, suggesting that treatment integrity would likely have been maintained in this study as well. On the other hand, those performance feedback studies were conducted prior to rigorous single case design standards and the intervention provided was not a standard protocol class-wide intervention. Further research should explore treatment integrity of a standard protocol class-wide intervention after performance feedback is withdrawn.

A third limitation of the study is the small student sample available for the t-tests used to detect differences between baseline and performance feedback growth rates. The multiple baseline design was primarily intended to examine the impact of performance

feedback on treatment integrity. For the t-tests, the samples sizes were sufficient to detect only medium and large effect sizes, but not small effect sizes.

Implications for practice

In accordance with previous studies on treatment integrity, this study indicates that evidence-based, class-wide, standard protocol interventions with prescriptive behaviors may not be implemented with integrity in schools. Researchers and school administrators should not assume and expect that all teachers will implement such interventions without some support. If higher treatment integrity is related to better student outcomes, then supporting teachers in maintaining treatment integrity is essential to bridging the research to practice gap in effectively delivering reading interventions.

Fortunately, intervention integrity can be improved through consultation. Direct observation and provision of performance feedback by a consultant or school staff member who is thoroughly trained on the intervention is an effective way to monitor and enhance implementation quality. As this study demonstrated, performance feedback can be effective for large scale, class-wide interventions, and for interventions implemented in urban schools districts with large percentages of at-risk students. Furthermore, this study and the one by Hagermoser Sanetti and colleages (2013) demonstrated that internal school staff can be trained to monitor and provide feedback on implementation, as an alternative to relying on outside research assistants. The consultants could be a school social worker or special education teacher (as was the case in Hagermoser Sanetti et al., 2003), RTI specialists (as was the case in this study), curriculum specialists, the school psychologist, or teachers that have successfully implemented an intervention with

integrity. As Vaughn and colleagues (2000) conveyed, school staff are typically more aware of the realities and needs of a particular school than researchers are. In addition, costs are likely lower when internal consultants or school staff are trained than when researchers or new outside consultants are used. Thus, performance feedback delivered by pre-existing school staff can be an effective and efficient way to improve the implementation of interventions in schools.

Future Research

This study is the first to use rigorous multiple baseline design to examine the impact of performance feedback on treatment integrity of a class-wide intervention. It is also the first multiple baseline performance feedback study to simultaneously analyze outcomes for groups of ELLs and native English speakers. With standards for single case design now available and effect sizes estimates continuously being developed, high quality multiple baseline studies examining the impact of performance feedback should be conducted with more class-wide, standard protocol, academic and behavioral interventions. In addition, since only one other multiple baseline study on performance feedback has used internal consultants (Hagermoser Sanetti, Fallon, & Collier-Meek, 2013), future studies on treatment integrity should examine the effectiveness and acceptability of having internal school staff deliver performance feedback.

Finally, further research on performance feedback for class-wide interventions should also include progress monitoring for a larger sample of students to increase power to detect improvements in student outcomes. To focus on examining the impact of performance feedback on growth rates, a group design study (e.g. randomized controlled

trial) can be used. Growth rates in classrooms implementing an intervention without performance feedback can be compared to growth rates in classrooms implementing an intervention with performance feedback. A group design with a larger number of classrooms would increase power to detect improved outcomes.

References

Begeny, C. B., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review, 40*(1), 23 – 38.

Boardman, A., Arguelles, M. E., Vaughn, S., Hughes, M. T., & Klingner, J. (2005). Special education teachers' views of research-based practices. *The Journal of Special Education, 39*(3), 168 – 180.

Christ, T. J., & Silberglitt, B. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128 – 133.

Cialdini, R. B. (2006). *Influence: the psychology of persuasion.* New York: Harper Business.

Cochrane, W. S. & Laux, J. M. (2008). A survey investigation school psychologists' measurement of treatment integrity in school-based interventions and their beliefs about its importance. *Psychology in the Schools, 45*(6), 499 – 507.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Delquadri, J., Greenwood, C. R., Stretton, K., & Hall, R. V. (1983). The peer tutoring game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children, 6*, 225 – 239.

DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of

mentoring programs for youth: A meta-analytic review. *American Journal of Community Psychology, 30*(2), 157–198.

Durlack, J. A., & Dupre, E. P. (2008). Implementation matters: A review on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41,* 327 – 350.

Fantuzzo, J. W., Polite, K., & Grayson, N. (1990). An evaluation of reciprocal peer tutoring across elementary school settings. *Journal of School Psychology, 28,* 309–323.

Fiala, C., & Sheridan, S. M. (2003). Parent involvement and reading: using curriculum-based measurement to assess the effects of paired reading. *Psychology in the Schools, 40*(6), 613 – 626.

French, J. R. P., & Raven, B. (1959). The bases of social power. In D. Cartwright and A. Zander (Eds) *Group dynamics*. New York: Harper and Row.

Fuchs, L.S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Fuchs, D., & Fuchs, L. S. (2001). One blueprint for bridging the gap: Project PROMISE: (Practitioners and researchers orchestrating model innovations to strengthen education). *Teacher Education and Special Education, 24*(4), 304 – 314.

Fuchs, L.S., Fuchs, D., Hamlett, C.L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27 – 48.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-Assisted Learning

Strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34*(1), 174 – 206.

Fuchs, D., Fuchs, L. S., Simmons, D. C., & Mathes, P. G. (2008). *Peer Assisted Learning Strategies: Reading Methods for Grades 2 – 6.*

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*(3), 157 – 171.

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools, 45*(6), 537 – 549.

Greenwood, C. R. (2009). Treatment integrity: revisiting some big ideas. *School Psychology Review, 38*(4), 547 – 553.

Greenwood, C. R., & Abbott, M. (2001). The research to practice gap in special education. *Teacher Education and Special Education, 24*(4), 276 – 289.

Greenwood, C. R., Arreaga-Mayer, C, Utley, C. A., Gavin, K. M., & Terry, B.J. (2001). Classwide peer tutoring learning management system: Applications with elementary-level English language learners. *Remedial and Special Education, 22*(1), 34 – 47.

Greenwood, C. R., Terry, B., Arreaga-Mayer, C., & Finney, D. (1992). The Classwide Peer Tutoring Program: Implementation factors that moderate students' achievement. *Journal of Applied Behavior Analysis, 25*(2), 101 – 116.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*(1), 37 – 50.

Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review, 22*(2), 254 – 272.

Gresham, F. M., & Kendall, G. K. (1987). School consultation research: methodological critique and furture research directions. *School Psychology Review, 16,* 306 – 316.

Guli, L. A. (2005). Evidence-based parent consultation with school-related outcomes. *School Psychology Quarterly, 20*(4), 455 – 472.

Hagermoser Sanetti, L. M., Fallon, L. M., & Collier-Meek, M. A. (2013). Increasing teacher treatment integrity through performance feedback provided by school personnel. *Psychology in the Schools, 50* (2).

Hagermoser Sanetti, L., M., Gritter, K. L., & Dobey, L. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review, 40*(1), 72 – 84.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 38*(4), 445 – 459.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2008). Treatment integrity in behavioral consultation: measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4*(1), 95 – 114.

Heartland Area Education Agency (2006). Grades 2-6 Reading PALS Implementation

Integrity Direct Observation Checklist. Retrieved from http://www.aea11.k12.ia.us/educators/idm/checkists.html.

Horner, R. H., Carr E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research design to identify evidence-based practice in special education. *Council for Exceptional Children, 71*(2), 165 – 179.

Kearns, D. M., Fuchs, D., McMaster, K. L., Saenz, L., Fuchs, L. S., … Yen, L. (2010). Factors contributing to teachers' sustained use of kindergarten peer- assisted learning strategies. *Journal of Research on Educational Effectiveness, 3,* 315 – 342.

Kratochwill, T. R., & Bergan, J. R. (1990). *Behavioral consultation in applied settings: An individual guide.* New York: Plenum Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., … Shadish, W. R. (2010). Single-case design technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L … Shadish, W. R. (2013).Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26 – 38.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development and Education, 52*, 33 – 42.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in

single- case research: Application examples. *Journal of School Psychology*, *49*, 301-321.

Maheady, L., Mallette, B., & Harper, G. F. (2006). Four classwide peer tutoring models: Similarities, differences, and implications for research and practice. *Reading & Writing Quarterly*, *22*, 65 – 89.

Mathes, P. G. & Babyak, A. E. (2001). The effects of peer-assisted literacy strategies for first-grade readers with and without additional mini-skills lessons. *Learning Disabilities Research & Practice, 16*(1), 28 – 44.

McMaster, K. L., Fuchs, D., & Fuchs, L. S. (2007). Promises and limitations of peer-assisted learning strategies in reading. *Learning Disabilities: A Contemporary Journal, 5*(2), 97 – 112.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*(4), 445 – 463.

Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review*, *27*(4), 613 – 627.

National Center for Education Statistics (2012). *The Condition of Education 2012* (NCES Report No. 2011-045). Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/pubs2012/2012045.pdf

National Center for Education Statistics (2011). *The nation's report card: reading 2011* (NCES Report No. 2012-457). Washington, DC: U. S. Department of Education.

Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf

Noell, G. H. (2010). Empirical and pragmatic issues in assessing and supporting intervention implementation in school. In G. Gimpel Peacock, R. A. Ervin, E. J. Daly, & K. W. Merrell (Eds), *Practical handbook of school psychology: Effective practices for the 21st century* (pp. 513 - 527). New York: Guilford Press.

Noell, G. H. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 323-342). Mahwah, NJ: Erlbaum.

Noell, G. H., & Gansle, K. A. (2006). Assuring the form has substance: Treatment plan implementation as the foundation of assessing response to intervention. *Assessment for Effective Intervention, 32*(1), 32 – 39.

Noell, G. H. & Witt, J. C. (1999). When does consultation lead to intervention implementation? Critical issues for research and practice. *The Journal of Special Education, 33*(1), 29 – 35.

Noell, G. H., Witt, J. C., Gilbertson, D. N., Ranier, D. D., & Freeland, J. T. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly, 12*(1), 77 – 88.

Noell, G. H., Witt, J. C., LaFleur, L. H., Mortenson, B. P., Ranier, D. D., & Levelle, J. (2000). Increasing intervention implementation in general education following consultation: A comparison of two follow-up strategies. *Journal of Applied*

Behavior Analysis, 33(3), 271 – 284.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., … Resetar, J. L. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies, *School Psychology Review, 34*(1), 87 – 106.

Parker, R. I., Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95 – 105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. J. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*(4), 194 – 204.

Parker, R. I., Vannest, K. J. & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional children, 75*(2), 135 – 150.

Pearson (2012). *Aimsweb Technical Manual*. Retrieved from http://www.aimsweb.com/

Pearson (2011). *Aimsweb default cutoff scores explained.* Retrieved from http://www.aimsweb.com/

Persampieri, M. Gortmaker, V., Daly III, E. J., Sheridan, S. M. & McCurdy, M. (2006). Promoting parent use of empirically supported reading interventions: Two experimental investigations of child outcomes. *Behavioral Interventions, 21*, 31 – 56.

Power, T. J., Blom-Hoffman, J., Clarke, A. T., Riley-Tillman, T. C., Kellerher, C., & Manz, P. (2005). Reconceptualizing intervention integrity: A partnership-based framework for linking research with practice. *Psychology in the Schools, 42*(5),

495 – 507.

Ransford, C. R., Greeberg, M. T., Domitrovich, C. E., Small, M., & Jacobson, L. (2009). The role of teachers' psychological experiences and perceptions of curriculum supports on the implementation of a social and emotional learning curriculum. *School Psychology Review, 38*(4), 510 – 532.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Raykov, T., & Marcoulides, G. A. (2013). *Basic statistics: An introduction with R.* Lanham, MD: Rowman and Littlefield Publishers, Inc.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95*(2), 240 – 257.

Saenz, L. M., Fuchs, L. S., & Fuchs, D. (2005). Peer-assisted learning strategies for English language learners with learning disabilities. *Exceptional Children, 71*(3), 231 – 247.

Sanetti, L. M., Kratochwill, T. R, & Long, A. C. L. (2013). Applying adult behavior change theory to support mediator-based intervention implementation. *School Psychology Quarterly, 28*(1), 47 – 62.

Shinn, M. R., Gleason, M. M., & Tindal, G. (1989). Varying the difficulty of testing materials: Implications for curriculum-based measurement. *The Journal of Special Education, 23*, 223 – 233.

Shinn, M. M. & Shinn, M. R. (2002a). Administration and scoring of reading MAZE for use in general outcome measurement. *Edformation Inc.*

Shinn, M. M. & Shinn, M. R. (2002b). Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement. *Pearson.*

Sindelar, P. T., & Brownell, M. T. (2001). Research to practice dissemination, scale, and context: we can do it, but can we afford it? *Teacher Education and Special Education*, *24*(4), 348 – 355.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510 – 550.

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of single case literature. *School Psychology Review, 41*(2), 160 – 175.

Stein, M. L., Berends, M. Fuchs, D., McMaster, K., Saenz, L., Yen. L., ... Compton, D. L, (2008). Scaling up and early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30*(4), 368 – 388.

Thomas, R. L. (1993). Cross-age and peer tutoring. *ERIC Digest.* Bloomington, IN: ERIC Clearinghouse on Reading and Communication Skills.

Tilly III, W. D., Niebling, B. C., and Rahn-Blakeslee, A. (2010). In G. G. Peacock, R. A. Ervin, E. J. Daly III, & K. W. Merrell (Eds) *Practical Handbook of School Psychology: Effective Practices for the 21st Century* (pp. 579-596). New York, NY: The Guildford Press.

U. S. Department of Education, Institute of Education Sciences (2011). What works clearinghouse: Procedures and Standards Handbook. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

U. S. Department of Education, Institute of Educational Sciences (2012a). What works clearinghouse intervention report, beginning reading: Peer assisted learning / literacy strategies. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_pals_050112.pdf

U. S. Department of Education, Institute of Educational Sciences (2012b). What works clearinghouse intervention report, adolescent Reading: Peer assisted learning strategies. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_pals_013112.pdf

U. S. Department of Education, Institute of Educational Sciences (2010). What works clearinghouse intervention report, English language learners: Peer assisted learning strategies. Retrieved from http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=366

Vadasy, P. F., Jenkins, J. R., Antil, L. R., Phillips, N. B., & Pool, K. (1997). The research to practice ball game: Classwide peer tutoring and teacher interest, implementation, and modifications. *Remedial and Special Education, 18*(3), 143 – 156.

Vaughn, S., Denton, C. A., & Fletcher, J. M. (2010). Why intensive interventions are necessary for students with severe reading difficulties. *Psychology in the Schools,*

47(5), 432 – 444.

Vaughn, S., Klingner, J., & Hughes, M. (2000). Sustainability of research-based practices. *Exceptional Children, 66*(2), 163 – 171.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541– 561.

Wickstrom, K. F., Jones, K. M., LaFleur, L. H., and Witt, J. C. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13*(2), 141 – 154.

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The effects of school-based intervention programs on aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology, 71*(1), 136–149.

Witt, J. C., Noell, G. H., LaFleur, L. H., & Mortenson, B. P. (1997). Teacher usage of interventions in general education: Measurement and analysis of the independent variable. *Journal of Applied Behavior Analysis, 30*(4), 693 – 696.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*(2), 203 – 214.

Figure 1
Logic Model

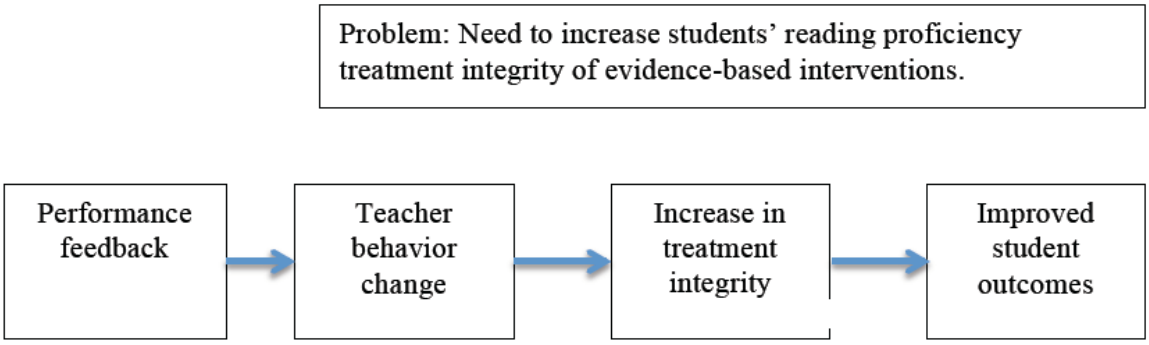

Figure 2
Treatment Integrity Checklist

Grades 2-6 Reading PALS
Vanderbilt University, Peabody College http://kc.vanderbilt.edu/pals/

Implementation Integrity Direct Observation Checklist

Teacher: _____ School: _____ Observer: _____

Student: _____ Grade: _____

Lesson #: _____ Start Time: _____ End Time: _____

Directions: During the observation, place a checkmark in the "+" (or "-") column for each step observed (or not observed). Tally the number of "+" and calculate integrity for each lesson part and overall integrity (see summary form at end of this sheet).

Note: If the step is not applicable, write N/A in the "+" column and do not include in the calculation of fidelity (for each part or overall total).

Part I: Classroom Arrangement/Set-up

+	-	Step	Checklist
		1	Student pairs are posted on bulletin board or transparency
		2	Teacher materials (e.g., overheads, timer) are organized
		3	Student materials (e.g., pencils, point sheets, question cards) are available or in folders
		4	Student books with marked page numbers or bookmarks are available

Number of +/4 = _____% Classroom Arrangement/Set-up Fidelity

Part II: Partner Reading
(Note: Observe at least two pairs)

Pair One		*Pair Two*			
+	-	+	-	**Step**	**Checklist**
				1	Reader 1 reads aloud from book for 5 minutes Start: ____ End: ____
				2	Reader 2 corrects mistakes using the correction procedure (e.g., "Stop. You missed that word. Can you figure it out? [waits approximately 4 seconds] Good. Read the sentence again." Or "That word is ___. What word? Good. Read the sentence again.")
				3	Reader 2 awards 1 point for each correctly read sentence
				4	Pairs switch jobs after 5 minutes.
				5	Reader 2 reads the same text for 5 minutes Start: ____ End: ____
				6	Reader 1 corrects mistakes using the correction procedure (e.g., "Stop. You missed that word. Can you figure it out? [waits approximately 4 seconds] Good. Read the sentence again." Or "That word is ___. What word? Good. Read the sentence again.")
				7	Reader 1 awards 1 point for each correctly read sentence
				8	Reader 2 retells the story for 1 minute (Gr. 2-3) or 2 minutes (Gr. 4-6) Retell prompts: "What happened first?" and "What happened next?" Start: ____ End: ___
				9	Students mark a total of 10 points for retelling story

Pair 1: Number of +/9 = _____% Partner Reading Fidelity
Pair 2: Number of +/9 = _____% Partner Reading Fidelity

Part III: Paragraph Shrinking
(Note: Observe at least two pairs)

Pair One		*Pair Two*			
+	-	+	-	Step	Checklist
				1	Reader 1 reads aloud from NEW text for 5 minutes Start: ____ End: ____
				2	Reader 1 names the most important "who" or "what" in the paragraph
				3	Reader 2 awards 1 point for correct answer
				4	Reader 1 states the most important thing about the "who" or "what"
				5	Reader 2 awards 1 point for correct answer
				6	Reader 1 states the main idea in 10 words or less
				7	Reader 2 awards 1 point for correct answer
				8	Reader 2 helps fix answers using the correction procedure (e.g., "That's not quite right. Skim the paragraph and try again.")
				9	Pairs switch jobs after 5 minutes
				10	Reader 2 reads aloud from NEW text for 5 minutes Start: ____ End: ____
				11	Reader 2 names the most important "who" or "what" in the paragraph
				12	Reader 1 awards 1 point for correct answer
				13	Reader 2 states the most important thing about the "who" or "what"
				14	Reader 1 awards 1 point for correct answer
				15	Reader 2 states the main idea in 10 words or less
				16	Reader 1 awards 1 point for correct answer
				17	Reader 1 helps fix answers using the correction procedure (e.g., "That's not quite right. Skim the paragraph and try again.")

Pair 1: Number of +/17 = _____% Paragraph Shrinking Fidelity
Pair 2: Number of +/17 = _____% Paragraph Shrinking Fidelity

Part IV: Prediction Relay
(Note: Observe at least two pairs)

Pair One		*Pair Two*			
+	-	+	-	**Step**	**Checklist**
				1	Reader 1 predicts what will happen in the text
				2	Reader 2 assigns 1 point for making a reasonable prediction
				3	Reader 1 reads a half page of NEW text
				4	Reader 2 assigns 1 point for reading half page
				5	Reader 2 asks Reader 1, "Did your prediction come true?"
				6	Reader 1 answers, "Yes," "no," or "don't know yet."
				7	Reader 2 assigns 1 point for Reader 1's response
				8	Reader 1 makes a new prediction (and process of predicting and assigning points continues until pairs switch roles)
				9	Pairs switch jobs after 5 minutes
				10	Reader 2 predicts what will happen in the text
				11	Reader 1 assigns 1 point for making a reasonable prediction
				12	Reader 2 reads a half page of NEW text
				13	Reader 1 assigns 1 point for reading half page
				14	Reader 1 asks Reader 2, "Did your prediction come true?"
				15	Reader 2 answers, "Yes," "no," or "don't know yet."
				16	Reader 1 assigns 1 point for Reader 2's response
				17	Reader 2 makes a new prediction (process of predicting and assigning points continues until time ends)

Pair 1: Number of +/17 = _____% Prediction Relay Fidelity
Pair 2: Number of +/17 = _____% Prediction Relay Fidelity

Part V: General Teacher Behaviors

+	-	Step	Checklist
		1	Most pairs (most =80%; in a class of 20, 8 of 10 pairs) actively follow along and are engaged in activities
		2	Teacher monitors most pairs (most =80%; in a class of 20, 8 of 10 pairs) throughout the PALS lesson
		3	Teacher awards extra points to individuals and/or large group for good PALS behaviors
		4	Provides positive feedback to individuals and/or large group
		5	Provides corrective feedback individuals and/or large group (as needed)
		6	Partner Reading lasts 10 minutes
		7	Story Retell lasts 1 minute (Grades 2-3) or 2 minutes (Grades 4-7)
		8	Paragraph Shrinking lasts 10 minutes
		9	Prediction Relay lasts 10 minutes *(Note: Prediction Relay is introduced in Week 5)*

Number of +/8 (or 9) = _____% General Teacher Behaviors Fidelity

Summary

Activity	Number of +	Total Number Possible	%
Classroom Arrangement/Set-Up		4	
Partner Reading Pair One		9	
Partner Reading Pair Two		9	
Paragraph Shrinking Pair One		17	
Paragraph Shrinking Pair Two		17	
Prediction Relay Pair One		17	
Prediction Relay Pair Two		17	
General Teacher Behaviors		8 or 9	
Overall Grade 2-6 Reading PALS Integrity		98 or 99	

Figure 3
Consultation Procedural Integrity Checklist

Week _____

	Observation 1 Date_____	Observation 2 Date _____
Provided feedback prior to today's PALS about previous observation. Including:		
▪ One strength or improvement.	_____	_____
▪ Corrective feedback on missed items corresponding to key PALS components.	_____	_____
▪ Asked teachers to remind entire class about the missed student items.	_____	_____
▪ Showed teacher TI graph up to previous observation.	_____	_____
▪ Completed TI checklist for today's PALS.	_____	_____
▪ Calculated today's percentage treatment integrity and added the data point to the TI graph.	_____	_____

Progress Monitoring
Date_____

- o Progress monitored 3 lowest ELL students and 3 lowest non-ELL students.
- o Showed teacher student progress monitoring graphs for each student.

Figure 4
Consultant Rating Profile

Consultant Rating Profile

Teacher: _____ Consultant: _____ Date:_____

Please rate your agreement with each of the following statements. For the first 7 items, please rate your agreement **for the period of time the RTI specialist provided feedback on PALS implementation.**

Statement:		Strongly Disagree		Slightly Disagree		Slightly Agree		Strongly Agree
1. The feedback the RTI Specialist provided was helpful.		1	2	3	4	5	6	7
2. The RTI Specialist listened to my concerns.		1	2	3	4	5	6	7
3. Communication with the RTI Specialist was timely and helpful.		1	2	3	4	5	6	7
4. The performance feedback process was a good use of my time.		1	2	3	4	5	6	7
5. I would choose to seek help from this RTI specialist again in the future.		1	2	3	4	5	6	7
6. The assessment and graphs of students' reading progress was useful and informative.		1	2	3	4	5	6	7
7. The graph showing % treatment integrity obtained in PALS implementation was useful.		1	2	3	4	5	6	7
8. PALS was implemented as planned.		1	2	3	4	5	6	7
9. PALS was effective.		1	2	3	4	5	6	7
10. I was satisfied with PALS' effectiveness.		1	2	3	4	5	6	7

Adapted from Noell et al., 2005

Figure 5
Impact of performance feedback on percent treatment integrity

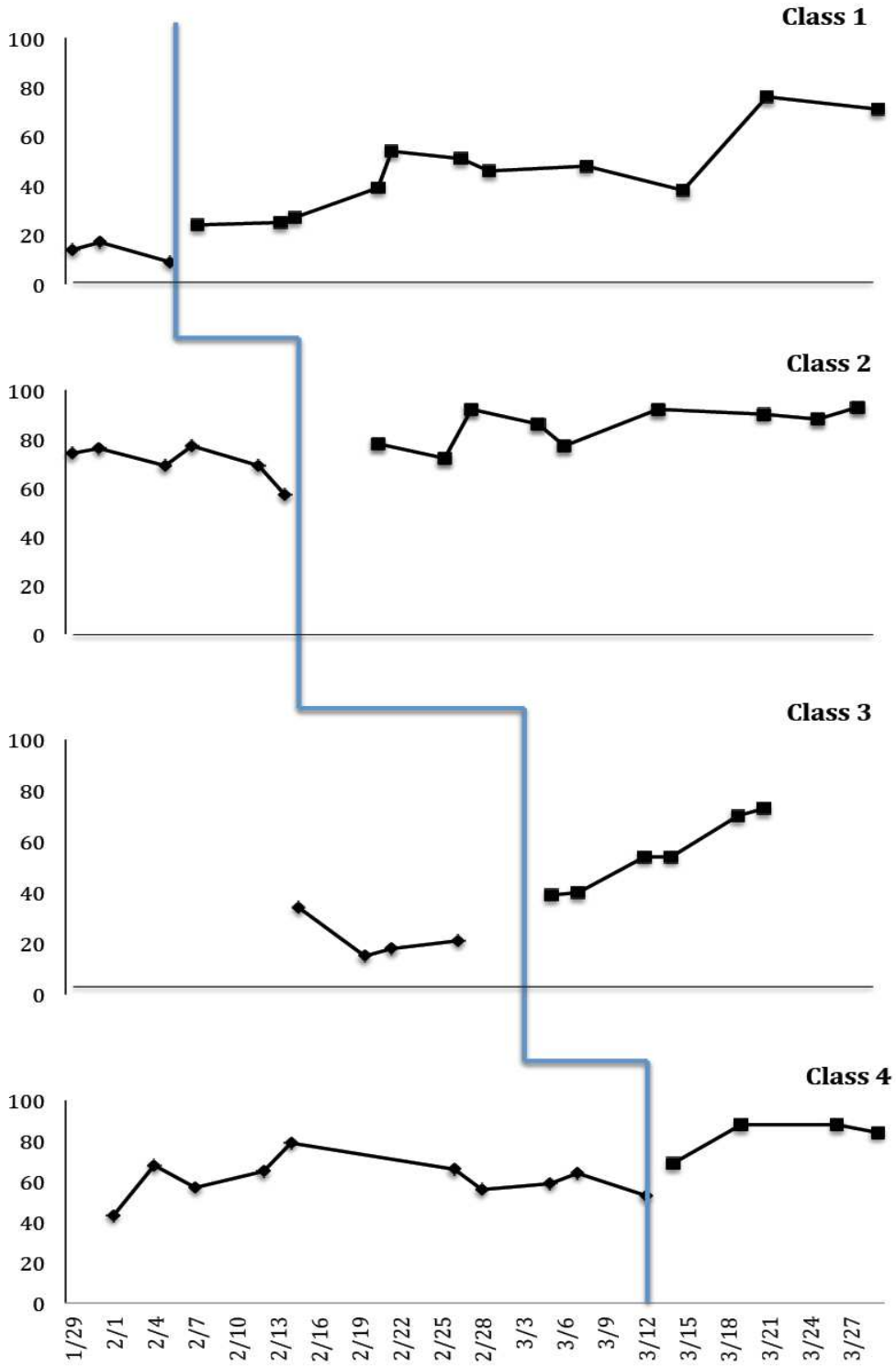

Figure 6

Impact of Performance Feedback on Oral Reading Fluency

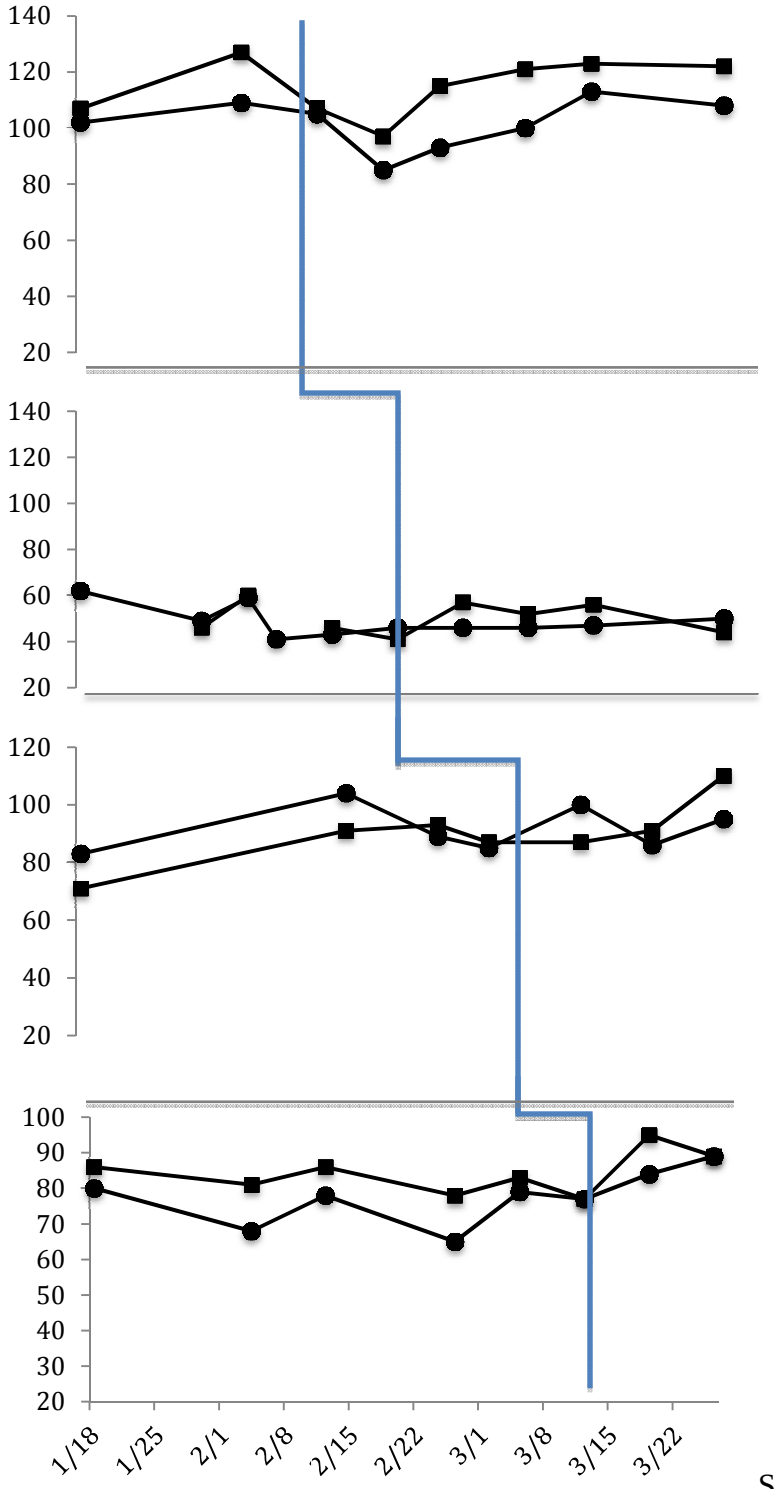

Table 1
Initial oral reading fluency scores

Classroom	Grade	Number of students in class	Average ORF before PALS	Percentile on Aimsweb National Norms
1	2	32	122 words per minute	~85th percentile
2	3	17	59 words per minute	~10th percentile
3	5	27	116 words per minute	~25th percentile
4	5	27	126 words per minute	~45th percentile

Table 2
2 x 2 Table for Calculating Φ

Overlap	Intervention	Baseline	Total
Higher	*a*	*b*	% baseline points
Lower	*c*	*d*	% intervention points
Total	% baseline points	% intervention points	100

Parker, Hagan-Burke, & Vannest, 2007

Table 3
Treatment integrity of PALS core components

Core Component	% of baseline observations in which component was implemented	% of PF observations in which component was implemented
All students are paired.	100	100
Students are reminded to sit side by side.	91	100
Higher reader reads first.	100	100
Students use question cards.	74	100
Students use correction cards	39	63
Use of the points system.	74	67
Partner reading done.	100	100
Partner reading timed correctly.	74	93
Retell done.	100	100
Retell timed correctly.	57	100
Paragraph shrinking done.	70	100
Paragraph shrinking timed correctly.	65	83
Prediction relay done.	87	83
Prediction relay timed correctly.	70	57
Teacher monitors pairs for 75% of the time.	91	97
Teacher provides corrective feedback	91	97
Teacher gives positive feedback or bonus points for desired tutoring/reading behaviors.	74	87

Adapted from Vadasy et al., 1997

Table 4
Change in treatment integrity

	Baseline Average	Performance Feedback Average	TI at Final Session
Classroom 1	13%	45%	71%
Classroom 2	70%	85%	93%
Classroom 3	22%	55%	73%
Classroom 4	61%	82%	84%
Average	42%	67%	80%

Table 5
IRD and R^2 Effect sizes

	IRD	95% CI	R^2	95% CI
Classroom 1	1.00	[.38 - 1.00]	.40	[.03 - .76]
Classroom 2	.28	[-.17 - .63]	.48	[.07 - .79]
Classroom 3	1.00	[.38 - 1.00]	.66	[.11 - .92]
Classroom 4	.65	[.11 - .87]	.50	[.04 - .83]
Average	.73		.51	

Table 6
Growth rates in words per week

	Mean (*SD*)	*t*	*p*
ELLs and native speakers combined			
Overall	.65 (.97)		
Baseline	2.05 (4.57)		
Performance feedback	2.86 (3.79)	.61	.55
ELLs only			
Overall	.49 (.79)		
Baseline	1.66 (4.50)		
Performance feedback	2.29 (3.43)	.36	.73
Native speakers only			
Overall	.82 (1.14)		
Baseline	2.47 (4.82)		
Performance feedback	3.47 (4.23)	.48	.64

Table 7
Consultant rating profile

Statement	Average rating (1 = strongly disagree, 7 = strongly agree)
1. The feedback the RTI Specialist provided was helpful.	6.75
2. The RTI Specialist listened to my concerns.	7
3. Communication with the RTI Specialist was timely and helpful.	6.5
4. The performance feedback process was a good use of my time.	6
5. I would choose to seek help from this RTI specialist again in the future.	5.5
6. The assessment and graphs of students' reading progress was useful and informative.	6.75
7. The graph showing % treatment integrity obtained in PALS implementation was useful.	5.5
8. PALS was implemented as planned.	6.75
9. PALS was effective.	5.5
10. I was satisfied with PALS' effectiveness.	5