INCREASING OBSERVATION AND FEEDBACK EFFICIENCY TO IMPROVE INSTRUCTIONAL QUALITY IN SMALL GROUP INTERVENTION SETTINGS

by

RONDA C. FRITZ

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2016

DISSERTATION APPROVAL PAGE

Student: Ronda C. Fritz

Title: Increasing Observation and Feedback Efficiency to Improve Instructional Quality in Small Group Intervention Settings

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Beth Harn	Chairperson
Brigid Flannery	Core Member
Audrey Lucero	Core Member
Gina Biancarosa	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2016

DISSERTATION ABSTRACT

Ronda C. Fritz

Doctor of Philosophy

Special Education and Clinical Sciences

September 2016

Title: Increasing Observation and Feedback Efficiency to Improve Instructional Quality in Small Group Intervention Settings

The current study investigated the reliability and validity of using short observations with an observation tool designed to measure implementation of small group interventions. Intervention lessons for eight instructional groups from two schools were video recorded for nine weeks, and post-test assessments of reading decoding were administered to 31 at-risk kindergarten students. Videos of intervention instruction from weeks two, five, and eight, each representing a phase in the intervention period, were used within this study for measuring implementation. Each video was divided into three ten-minute segments representing the beginning, middle, and end of each intervention lesson. Video segments were coded for implementation using the Quality of Intervention Delivery and Receipt tool (QIDR; Harn, Forbes-Spear, Fritz, & Berg, 2012). Overall, the results of this study indicate that a) reliability can be achieved when using 10-minute observations, b) QIDR scores obtained from 10-minute segments are strongly correlated with scores obtained from full-length observations, c) there is no statistical difference in scores obtained from full-length observations and those obtained in 10-minute segments, and d) QIDR scores obtained from both full-length and 10-minute segments accounted for group differences in student outcomes, with lesson segments obtained from the end of

lessons accounting for the most variance. Implications for research and practice are discussed, including the importance of thorough training and calibration to maintain reliability, as well as the feasibility and utility of providing frequent observation and feedback through shorter observations.

CURRICULUM VITAE

NAME OF AUTHOR: Ronda C. Fritz

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Boise State University, Boise, ID

DEGREES AWARDED:

Doctor of Philosophy, Special Education, 2016, University of Oregon
Master of Arts, Reading Education, 2001, Boise State University
Bachelor of Arts, Elementary Education, 1992, Boise State University

AREAS OF SPECIAL INTEREST:

Teacher Preparation; Early Reading Intervention; Student Engagement
Reading Methods
Elementary Teacher Education

PROFESSIONAL EXPERIENCE:

Assistant Professor of Education, Eastern Oregon University, La Grande, OR; September 2014-Present

Title I Reading Specialist, North Powder Charter School, North Powder, OR; August 2004-July 2011

$4^{th}/5^{th}$ Grade Classroom Teacher, North Powder School District, North Powder, OR; August 1996-June 1998

Middle School Language Arts/Mathematics Teacher, North Powder School District, North Powder, OR; August 1996-June 1998

Title I Reading Teacher, North Powder School District, North Powder, OR August 1993-June 1996

Kindergarten/Middle School Mathematics Teacher, Ukiah School District, Ukiah, Oregon; August 1992-June 1993

GRANTS, AWARDS, AND HONORS:

BASES Leadership Grant, University of Oregon, 2011-2016

William N. & Patsy A. Wilber Scholarship, University of Oregon, 2013-2014

Sammie Barker McCormack Scholarship, University of Oregon, 2013-2014

Dynamic Measurement Group Scholarship, University of Oregon, 2014-2015

PUBLICATIONS:

Harn, B., Basaraba, D., Chard, D., & Fritz, R. (2015). The Impact of Schoolwide Prevention Efforts: Lessons Learned from Implementing Independent Academic and Behavior Support Systems. *Learning Disabilities: A ContemporaryJournal*, *13*(1), 3-20

Fritz, R., (2014). Book review of: *Fingon, Joan C. & Ulanoff, Sharon H. (2012). Learning from culturally and linguistically diverse classrooms: Using inquiry to inform practice.* Bilingual Research Journal (37:1). **DOI:** 10.1080/15235882.2014.893269

Harn, B. A., Fritz, R., & Berg, T. (2014). How do we deliver high quality literacy and reading instruction in inclusive schools? In McLeskey, J. Waldron, N., Spooner, F., & Algozzine, B. (Eds.) *Handbook of Research and Practice for Effective Inclusive Schools*.

Fritz, R. (2001). *Accelerated Reader: A valuable tool for increasing reading achievement and motivation of at-risk fourth and fifth graders?* Submitted to fulfill requirements of Master's Thesis, Boise State University

ACKNOWLEDGMENTS

I wish to express sincere appreciation to my doctoral committee, Dr. Audrey Lucero, Dr. Brigid Flannery, Dr. Gina Biancarosa, and Dr. Beth Harn, for providing support throughout my doctoral program, and especially through the dissertation process. I want to especially thank, my advisor and dissertation committee chair, Dr. Beth Harn, for providing endless hours of support and feedback throughout my doctoral program. Thank you for knowing exactly how much encouragement and support I needed at just the right times. In addition, I would like to thank Tricia Berg, Tiffany Beattie, Manuel Monzalve, Kendra Carmen, Samantha Fritz, and Christine Aldrich for their endless dedication as part of my coding team, and Dr. Lina Shanley for your expertise and support during the analysis phase.

I would also like to thank Dr. James Sinclair, Dr. Allison Baker-Wilson, Dr. Ruby Batz, Dr. Kara Hirano, and Tricia Berg for making this doctoral journey bearable and even sometimes enjoyable! I also want to thank my former students and colleagues at North Powder Charter Elementary School for providing me with inspiration and encouragement to explore questions and answers for the benefit all of our students. Finally, I want to thank my husband, Shane, who endured countless hours in the car for visits, listened during endless hours of phone calls, and provided unending support throughout.

Dedicated to my son, my husband, and my father, who all provided the inspiration for my further studies; and to my daughter and mother for being my biggest cheerleaders. I love you all and appreciate your support and encouragement along the way.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Without adequate skill in reading, a child is likely to face many obstacles throughout education and life. Reading is a skill of profound social significance because it opens the door for subsequent education, which in turn expands opportunities for greater employment, enrichment, and entertainment (Saunders, 2011). During the past decades, great strides have been made to ensure the success of all children in early reading. There is substantial research documenting the effectiveness of many interventions in reducing the number of children with long-term reading difficulties (Gersten, Vaughn, Deshler, & Schiller, 1997; Simmons et al., 2011; Swanson, 1999), however, some students remain poor readers even after receiving highly intensive interventions (Denton, Fletcher, Anthony, & Francis, 2006).

There may be multiple factors that influence a child's responsiveness to high-quality, evidence-based interventions, including neurological, biological, and environmental factors (Shaywitz, 2008; Wolf, 2007). While these factors may be relevant, they focus solely on the student, and aren't readily malleable or changeable by educational personnel. One factor that is modifiable, and has received increased attention recently, is quality of instructional delivery. While most of the focus on this group of non-responders has focused on specific student characteristics (e.g., language status, ethnicity; Shaywitz, 2008; Torgesen, 2002), recent efforts have documented variability of instructional delivery and its impact on learning even when using evidenced-based programs (Cook & Odom, 2013).

Many have documented that an effective teacher is an important factor in a child's achievement (Chetty, Friedman, & Rockoff, 2011; Darling-Hammond, 2010; Hanushek & Rivkin, 2010; Hanushek et al., 2010). However, it is not uncommon for the students most at-risk to receive supplemental intervention from personnel with no formalized training, such as educational assistants (Causton-Theoharis, Doyle, Giangreco, & Vadasy, 2007). This lack of training and support may impact instructional quality and be at least partially responsible for non-response of some students for whom reading is most difficult.

Fixsen, Blase, Metz, & Van Dyke (2013) posit that improved outcomes can only be achieved when effective interventions are coupled with effective implementation. The authors also contend that effective implementation is obtained when adequate pre-service and in-service training, coaching, and performance assessment are provided. The reality of providing these types of supports to improve instruction in school settings is often far from this ideal. Coaching and supervision of the interventionist may be sparse in many school settings due to limited resources and/or an erroneous belief that evidence-based programs are "plug and play" and don't require preparation or resources for follow-up support (Fixsen et al., 2005 as cited in Fixsen, et al., 2013). There are two major challenges to providing this type of implementation support in schools: 1) identifying tools for measuring instructional quality for interventions, and 2) having time to complete the instructional evaluations. In the next sections, each of these challenges will be explored in more detail.

Tools for Measuring Quality

Studies focused on measuring the effectiveness of interventions have placed emphasis on evaluating implementation of the specific practice, or treatment fidelity, as part of a research project (Harn, Parisi, & Stoolmiller, 2013). Generally, treatment fidelity refers to the degree to which a treatment or intervention is delivered as intended (Yeaton & Sechrest, 1981). Measurement of treatment fidelity in educational research is often focused on structural and process fidelity (Gersten, Fuchs, et al., 2005; Odom, 2008). Structural fidelity refers to adherence to the central components of an intervention, dosage, and intervention completion (Durlak & DuPre, 2008; Gersten, Fuchs, et al., 2005) and is usually measured through direct observation or self-report by interventionists (Harn, et al., 2013). Process fidelity refers to the quality of intervention delivery and student-teacher interactions (Justice, Mashburn, Hamre, & Pianta, 2008). Some researchers have suggested that process fidelity is more difficult to define and measure, but may be more directly related to student outcomes than structural fidelity (Gersten, Fuchs, et al., 2005; Mowbray, Holter, Teauge, & Bybee, 2003). Holdheide, Browder, Warren, Buzick, & Jones (2012) stress the importance of measuring process components within school-level implementation because the goal in schools is to improve instructional delivery and quality, rather than to document intervention fidelity/adherence.

The tools being used to evaluate instructional quality have focused primarily on evaluating quality within general education classrooms (e.g., Cameron, Connor, & Morrison, 2005; Kane & Staiger, 2012; Pianta, Cox, Taylor, & Early, 2013) and have not focused on tools that can be used to formatively evaluate quality over time (e.g., Hagan-

Burke et al., 2013; Johnson & Semmelroth, 2013). Educators and educational researchers acknowledge that evaluation of instructional quality within general education is necessary to ensure that all children are receiving high quality instruction. Although other means of measuring quality have been used such as value-added models (VAMs), teacher self-report, and student evaluations (Kane & Staiger, 2012), observation remains one of the most widely used methods for gaining a more direct measure of classroom interactions that may be impacting student outcomes (Chomat-Mooney et al., 2008). As a result, multiple observation tools have been developed for this purpose and have been found reliable and valid for observation and evaluation in the general education setting (e.g., Danielson, 1996; Fish & Dane, 2000; La Paro, Pianta, & Stuhlman, 2004; Maxwell, Mcwilliam, Hemmeter, Ault, & Schuster, 2001; Waxman, Huang, Anderson, & Weinstein, 1997). These tools are designed around a definition of quality as it applies in a whole-class setting and may not accurately measure quality of instruction in a small-group intervention setting (Johnson & Semmelroth, 2013).

Measuring instructional quality in general education. In an attempt to differentiate instruction for a diverse group of learners, teachers in general education classrooms may need to use a variety of instructional approaches to adapt to the needs of a wide variety of learners (Yopp & Yopp, 2000). Therefore, the tools designed to measure instructional quality in this context often measure a broad sampling of teacher-student interactions and classroom environment factors. For instance, Pianta, La Paro, and Hamre (2008) developed the Classroom Assessment Scoring System (CLASS), which measures multiple instructional dimensions including emotional support, classroom management, and instructional support, using ten different dimensions (e.g.,

positive and negative climate, teacher sensitivity, behavior management, concept development, language modeling). Danielson (1996) also developed a system for measuring classroom quality called the Framework for Teaching. This system includes 22 subscales for measuring planning and preparation, classroom environment, instruction, and professional responsibilities. Research regarding these and other observation tools in general education has provided insight into effective ways to measure instructional quality (e.g., Danielson, 2011; Darling-Hammond, 2010; Kane & Staiger, 2012), however, there is a need to shift our focus from examining not only what quality looks like in general education, but to understand how to measure quality of instruction designed for our most at-risk students receiving intervention supports (Johnson & Semmelroth, 2012; Semmelroth & Johnson, 2013).

Measuring instructional quality in intervention. A tool designed for use in intervention settings must reflect the differences between instructional quality in general education and intervention settings. The definition of instructional quality is apt to be considerably different in these two contexts. While differentiated instruction in a general education classroom calls for varying instructional approaches adapted to the needs of diverse students (Hall, Vue, Strangman, & Meyer, 2014), intervention settings are likely to need much more specific approaches to meeting the needs of individual students (Zigmond & Kloo, 2011).

Instruction for intervention efforts is designed for the purpose of accelerating learning focusing on individualizing instruction (Justice, 2006). To maximize learning, intervention efforts are designed and implemented very differently than in general education. Intervention is delivered in small groups, under a more focused time constraint

(i.e., 20 to 30 minutes in length), and must be intensive, focused, and explicit (Foorman & Torgesen, 2001; Torgesen et al., 1999). Specific, systematic, direct instruction coupled with explicit strategy instruction has shown positive effect sizes in student achievement (Gersten et al., 1997; Swanson, 1999). This type of instruction is often focused on basic skills and may not lend itself to some of the interactions measured by tools designed for general education (Forbes-Spear, 2014). For instance, within the CLASS tool (Pianta et al., 2008) one of the variables measured is concept development. While this may be an important construct in general education instructional contexts, this type of interaction may not be essential within a curriculum that is concentrated on basic skill development (Semmelroth & Johnson, 2013). The short duration and intensive, specific focus of an intervention session limits the variety of interactions, which should be reflected in the type of tool used to measure quality.

Tools designed for use in intervention contexts must be specifically measuring skills essential for accelerated learning. Johnson and Semmelroth (2012) have begun to explore this idea through development of the Recognizing Effective Special Education Teachers (RESET) tool for measuring instructional quality within Special Education settings. This tool reflects some of the differences in what is considered "instructional quality" between general education and intervention contexts. The RESET tool is based on the instructional domain of the Danielson (1996) Framework for Teaching (FFT), but was adapted to clearly delineate instructional components necessary for delivering evidence-based practices to students with disabilities, rather than the more constructivist approach to instructional delivery reflected in the FFT (Semmelroth, 2013). The RESET observation tool contains between 28 and 67 items (depending on the number of

instructional practices being observed) and consists of three main parts: Lesson Overview (introduction), Lesson Components (instructional practices), and Lesson Summary (conclusion; Johnson & Semmelroth, 2012). The tool is designed to be used with video-taped lesson footage to provide feedback to Special Education teachers.

The development of the RESET tool for measuring the quality of interventions delivered by special education teachers moves the field of education closer to providing a solution for measuring instructional quality in alternative settings. However, as was mentioned previously, intervention at the tier two and tier three levels within a response to intervention (RtI) framework is commonly delivered by educational assistants (Causton-Theoharis et al., 2007). While the RESET tool provides a promising avenue for evaluating special education teachers, it does not necessarily provide a tool that can be used in any intervention setting (e.g., general education, special education, or Title I classroom) with any interventionist (e.g., instructional assistants, volunteers, or general education teachers). Although the tool was designed with an underlying intent to improve Special Education instruction, the length of the tool limits the efficiency necessary to provide frequent formative feedback to improve instruction.

Maximizing Time: Can We Measure Quality More Efficiently to Provide Regular Feedback?

Maximizing the efficiency of measurement of instructional quality is especially important when considering the importance of frequent formative assessment to improve instruction. While research on using a formative evaluation approach to the timely measurement of instructional quality is relatively uncommon, the use of formative assessment to measure student achievement with the purpose of determining whether or

not instruction is producing desired outcomes is frequently used (Stecker, Fuchs, & Fuchs, 2005). The heightened accountability in education of recent years has put an emphasis on teachers using formative assessments such as curriculum-based measures (CBMs) to determine if instruction is producing desired outcomes with students. The use of data-based decision-making has been shown to improve student achievement as teachers are attending more carefully to classroom-level data and modifying instruction as needed to produce desired results (Fuchs, Deno, & Mirkin, 1984; Stecker et al., 2005). Potentially, taking a similar, responsive approach to evaluating instructional quality would have the same benefit of improving student outcomes by focusing on the teacher.

Assessment of instructional quality must share other characteristics with CBMs in order to be effective and useful. CBMs are designed to be not only efficient, but sensitive enough to measure student growth across a short period of time (Good, Gruba, & Kaminski, 2002; Stecker, Lembke, & Foegen, 2008). In the same way, tools for formative assessment of instructional quality must also be efficient and sensitive in order to be effective and useful.

The efficiency of the tool is essential because multiple researchers have found that frequent observation and feedback produces greater achievement gains (Chomat-Mooney et al., 2008). To provide this level of support in school settings, tools that can provide targeted, specific feedback to improve instructional practice are necessary (Cook & Odom, 2013; Feng, Figlio, & Sass, 2010; Goe, Biggers, & Croft, 2012; Greenwood, Horton, & Utley, 2002; Kretlow & Bartholomew, 2010). Existing tools often focus on a multitude of variables which typically requires a longer duration for observation (whole class/intervention period), which may negatively impact the ability to provide timely,

frequent, and targeted feedback. It may be possible to provide an even more intense focus on fewer essential skills to maximize the efficiency of the tool, making observations of full lessons unnecessary.

Developing a more concise tool for evaluating instructional quality in an intervention context may allow for a shorter observation period to enable coaches and supervisors to conduct more frequent observations. Current tools designed for use in general education settings not only measure multiple aspects of quality that may not be applicable in intervention settings, these tools also require observation periods that are much longer than may be necessary in an intervention setting. Standard observation protocol using the CLASS (Pianta et al., 2008) requires at least four 30-minute cycles, for a total of two hours of observation time, to obtain reliable and valid scores of quality. Although the Framework for Teaching (Danielson, 1996) does not specify a time frame for observation, an entire lesson is required to measure all components of the tool, which is likely to mean no less than 20 minutes of observation. Even the RESET tool, designed specifically for intervention (Johnson & Semmelroth, 2012), requires the duration of an entire lesson (i.e., at least 15 minutes) to obtain information about all of the components within the tool.

The reality of providing numerous opportunities for feedback to interventionists is prohibitive as current tools are too lengthy or not focused on the facets of instructional quality specific and most critical to the intervention context. By further delineating the specific skills in intervention delivery that are responsible for improved student outcomes, shorter observation tools and observation periods may be possible, which would permit more frequent observation and feedback opportunities to improve overall

instructional quality. Given that administrators and coaches responsible for evaluating interventionists often have limited availability to conduct observations, provide feedback, and follow up to ensure improvement in instruction (Knight, 2007), efficient tools specifically designed for formative evaluation in intervention settings may be more practical than current tools. Greater clarification is needed to determine the essential features of quality intervention delivery as well as determining a sufficient amount of time needed to capture instructional quality.

Within the context of general education, one research team has begun to address the need for a more concise measurement tool to ensure more frequent and timely feedback. Gargani and Strong (2014) have developed a tool called the Rapid Assessment of Teacher Effectiveness (RATE). The premise behind this tool is to provide a means to identify successful teachers better, faster, and cheaper than current observation tools provide. The RATE tool has only ten items addressing general classroom practices (e.g., lesson objective, multiple delivery mechanisms, providing examples/non-examples, pacing) each rated on a scale of one to three, with three being the highest level of implementation. The RATE tool was specifically designed to predict the ability of teachers to raise the achievement of their students, using shorter segments of instruction, fewer observations, and less training (Gargani & Strong, 2014). Multiple studies of predictive validity have shown that the development of this tool has been successful in predicting teachers whose students will have the best achievement outcomes after only one twenty-minute observation (Strong, 2011). While this tool has been effective at identifying effective and non-effective teachers, the concise nature of the tool was not

intended to provide enough information to be used for formative assessment for improving instruction (Gargani & Strong, 2014).

Another research team has begun to explore the use of concise measurement of intervention implementation by using an observation tool called *Snippets* to study teachers' use of specific comprehension supports in a general education, whole-classroom setting (Pratt & Logan, 2014). In one study, two six-minute segments of a 90-minute observation video documenting the 90-minute reading block, were coded using the *Snippets* tool. Within the 90-minute reading block, teachers were instructed to use a supplemental curriculum called *Let's Know* (Language and Reading Research Consortium, in press) for 30 minutes and continue with regular language and reading activities for the remaining 60 minutes. One of the six-minute segments was extracted from the 30-minute *Let's Know* lesson, while the other was taken from the remaining 60 minutes of language and reading instruction. The *Snippets* tool was designed to measure very discrete language-focused supports related to comprehension development in primary-grade children. The focused nature of this tool allowed for reliable measures of the use of comprehension supports within this short time frame of six-minute observations. The tool was able to reliably measure significant differences between comprehension supports utilized during *Let's Know* instruction compared to Language Arts instruction outside of the *Let's Know* curriculum. Kappa was calculated indicating 86% reliability with a second observer for 14% of the coded segments.

Time is precious, particularly for students already performing below expectations. In the same way that the *Snippets* tool is specifically focused on one facet of reading instruction, a tool that focuses only on the most critical instructional practices for

improving student outcomes within intervention would give educators the ability to formatively evaluate intervention to improve outcomes. A tool that can allow for relatively short observation periods (i.e., 5-10 minutes) that can be repeated frequently (i.e., 3-4 times per month) could allow for a more responsive observational cycle and provide expedited improvement of instruction. An evaluation and support system that could provide this level of ongoing feedback for improving instructional delivery could ensure that student learning is accelerated and long-term outcomes for our most at-risk students are improved.

Purpose of the Study

Harn, Forbes-Spear, Fritz, and Berg (2012) developed the Quality of Instructional Delivery and Response (QIDR) tool to measure quality within small group intervention. The instructional elements in this tool have been shown through previous research to be important components of quality instruction in the intervention setting. The tool was originally designed to determine the relationship between instructional quality and student outcomes in early reading. Preliminary evidence indicates that the tool can be used reliably and that scores obtained are predictive of academic outcomes (Forbes-Spear, 2014). Using videos capturing small-group intervention instruction, the tool has been found to be reliable when measuring quality of a full-length lesson of 20-30 minutes (Harn, Forbes-Spear, Fritz, Berg, & Basaraba, 2014; Forbes-Spear, 2014).

The purpose of this study is to determine the relationship between shorter segments of intervention and the overall intervention session. The study is designed to determine if one can reliably measure instructional quality more efficiently by comparing a sub-sample of the intervention time (i.e., 10 minutes) to the overall intervention (i.e.,

25-30 minutes). In addition, this study examined if a specific period of time in the intervention is more related to the overall intervention delivery (e.g., beginning, middle, or end). These findings could assist schools in utilizing their supervisory personnel more efficiently which could maximize time to allow for a responsive observational cycle that would improve instructional quality for those students who need high-quality instruction most. Investigation of these issues was guided by the following research questions:

Research Questions

1) Can adequate inter-rater reliability (IRR) be obtained after observing 10 minutes of full-length intervention lessons?
2) Using the QIDR, what is the relationship between scores obtained watching the full lesson versus sampling ten minutes of the lesson?
3) To what extent does the relationship between QIDR ratings obtained watching the full lesson, versus sampling ten minutes of the lesson, depend on time segment of the lesson (i.e., beginning, middle, end) or on phase within the intervention (i.e., 2nd week, 5th week, 8th week)? In other words, are correlations between the ratings systematically stronger or weaker based on time segment or intervention phase?
4) Which QIDR ratings (full lesson vs. 10-minute sample; beginning, middle, end; intervention phase) account for the most variance in student outcomes?

CHAPTER II

LITERATURE REVIEW

One of the most influential factors in determining student achievement is the quality of their teacher and, specifically, the quality of instruction received (Darling-Hammond, 2010; Kane, Staiger, & McCaffrey, 2012). As a result, recent national policy has placed renewed emphasis on developing systems to evaluate teacher and instructional quality for the purpose of ensuring high-quality teachers, and improving instruction to maximize student outcomes (McGuinn, 2012; National Council on Teacher Quality, 2012). The implementation of these new policies has proven to bring about many challenges. Determining what constitutes a quality teacher, quality instruction, and the best way to measure that quality, are ongoing struggles in the field of education (Goe, Bell, & Little, 2008; Johnson & Semmelroth, 2012). The definition of quality instruction and what aspects of instruction most impact student outcomes continues to be pursued through research (e.g., Cameron et al., 2005; Carlisle, Kelcey, Berebitsky, & Phelps, 2011; Foorman & Torgesen, 2001; Gargani & Strong, 2014; Gersten, Baker, Haager, & Graves, 2005; Hagan-Burke et al., 2013; Pratt & Logan, 2014). The issue of measurement of instructional quality is further complicated by the variation in instructional contexts and the differences in instructional expectations for each of these settings (i.e., general vs. special education; Foorman & Torgesen, 2001; Zigmond & Kloo, 2011), as well as the resource-intensive nature of measurement when using current observation tools (Gargani & Strong, 2014).

Chapter two begins with a brief history of the use of classroom observation tools for research and school-based purposes, followed by an overview of the most commonly

used observation tools used to measure instructional quality in general education classrooms. For each tool, the specific purpose(s), content, and training and observation time requirements will be reviewed. Next, the literature review will focus on observation tools designed for the purpose of measuring instructional quality in alternative settings (i.e., intervention and special education), and the ways in which these tools, can and do, differ from those designed for general education. Next, a discussion will present recent research investigating the use of shorter observation periods to measure instructional quality and teacher effectiveness. Finally, the chapter concludes with a discussion of the need to determine more efficient and valid ways to evaluate instructional quality to support more effective intervention delivery in school settings.

Classroom Observation

Classroom observation has become a common component in the measurement of instructional quality and teacher effectiveness, both as an element of evaluation in applied settings, as well as for purposes of research (Chomat-mooney et al., 2008; Goe et al., 2012; Pianta, Mashburn, Downer, Hamre, & Justice, 2008; Semmelroth & Johnson, 2013). In applied settings, observation is sometimes used as part of high-stakes employment decisions (i.e., value-added measurement, raises, termination), but has also been found to be helpful for providing administrators with information that can guide professional development of teachers (Goe & Croft, 2009; Pianta, Mashburn, et al., 2008). Observation for the purpose of informing professional development is arguably the most important use, and standardized observation systems can provide a means for systematically determining needs for professional development for each teacher and school (Danielson, 2011; Pianta, 2003). The next section will highlight the historical

context of observation research before delving into specific tools designed for classroom observation.

History of Classroom Observation

Classroom observation has been a part of educational research for over forty years (Gage & Needels, 1989). Much of the initial research followed a process-product approach meaning that researchers were trying to identify which teacher processes (i.e., instruction and interactions) produced best student learning. This was also an attempt to delineate what made effective and ineffective teachers so that effective teaching could be emulated across classrooms (Brophy & Good, 1986; Brophy, 1986). This research was often quantitative in nature and focused on frequency counts of discrete classroom behaviors such as number of pages of curriculum presented, time allocated for instruction, or classroom management behaviors (Brophy & Good, 1986). This method of observation brought about various criticisms including that there was too much emphasis on discrete teacher behaviors as "causes" and student achievement as "effects," with no acknowledgement of various other classroom factors that might affect student achievement, including the reciprocal effect of teacher-student interactions (Gage & Needels, 1989; Gage, 1989). In addition to criticism regarding the content of observations, methodological concerns were also raised. Some believed that there was too much reliance on correlational research and advocated for an experimental approach to determine causality (Macmillan & Garrison, 1984).

As a result of these criticisms, researchers in the 1990s began to avoid use of quantitative methods and instead employed more qualitative approaches to observation (Chomat-Mooney et al., 2008; Gudmundsdottir, 1997). These methods provided rich

descriptions of complex classroom interactions and provided an avenue for creating hypotheses on what constituted high-quality classrooms, but findings were difficult to generalize and did little to provide definitive understanding of which teacher-student interactions allowed for the greatest student achievement (Chomat-Mooney, et al., 2008).

In more recent years, the establishment of the Institute for Education Sciences, through the Education Sciences Reform Act of 2002, has increased emphasis on providing rigorous scientific evidence for educational practices through increased availability of research funds to empirically study multiple aspects of education, including validated and standardized observation systems (Chomat-Mooney, et al., 2008). In addition, private research entities such as the William T. Grant, Spencer, and Bill and Melinda Gates Foundations have also provided funding for research on tools and systems for ensuring high-quality instruction to improve student outcomes (Chomat-Mooney, et al., 2008). Increased availability of funding has allowed various researchers to increase efforts to develop valid observation tools for measuring instructional quality (e.g., Cameron et al., 2005; NICHD, 2003; Pianta, Belsky, Houts, & Morrison, 2007).

Early observation systems were developed to account for multiple classroom features and interactions and were focused on early childhood settings. These included such observational tools as the Early Childhood Environment Rating Scale (ECERS; Harms & Clifford, 1980) and the Observational Record of the Caregiving Environment (ORCE; NICHD Early Child Care Research Network (ECCRN), 1996). Both of these measures were developed to measure the interactions of child care providers with children, as well as the overall quality of childcare settings. The revised edition of the ECERS (ECERS-R; Harms, Clifford, & Cryer, 1998) is still widely used as a global

measure of quality (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005), including teacher-student interactions, but has been found to have a greater focus on classroom environment than on interactions between teachers and students (Sammons et al., 2002). In contrast, the ORCE was developed to specifically measure interactions between the caregiver and individual children (NICHD ECCRN, 1996). The ORCE has been used in a large-scale longitudinal study to determine how various aspects of childcare quality impacted later student outcomes. One of the major findings of this study was that teacher's use of language (e.g., asking questions and responding to children's talking) was linked to better cognitive and language development (National Institute of Child Health and Human Development Early Child Care Research Network, 2000; NICHHD SECCYD).

An upward extension of the ORCE was later developed to measure quality in kindergarten and was called the Classroom Observation System for Kindergarten (COS-K; (National Center for Early Development and Learning [NCEDL], 1997) and was later adapted for first (COS-1; NICHD Early Child Care Research Network, 2002), and third and fifth grades (COS-3/5; NICHD Early Child Care Research Network, 2004). Both the ORCE and COS measure classroom features found to be related to students' academic and social development through time-sampling of discrete behaviors, coupled with more global rating scales, to capture quality of teacher-student interactions (Hamre & Pianta, 2005; NICHD Early Child Care Research Network, 2002b; Pianta, Belsky, Houts, Morrison, & National Institute of Child Health and Human Development Early Child Care Research Network, 2007; Rimm-Kaufman et al., 2002). The discrete behaviors were recorded during 10-minute periods of 30-second observation intervals and included

measures of setting and activities (e.g., teacher-directed activity, individual activity, unstructured activity, recess) and teacher behaviors (e.g., read-alouds, teacher-child interactions, teacher affect, and teaching of social and academic skills). Scoring of discrete behaviors was followed by global measures of classroom quality based on observations outside of the time-sampling of behaviors. Global ratings consisted of ratings of classroom dimensions such as overcontrol/intrusiveness, emotional climate, classroom management, literacy instruction, feedback, and child behavior. These dimensions were rated on a seven point scale, ranging from adequate to excellent (Pianta et al., 2002). The COS was a precursor to the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, et al., 2008) which will be discussed in detail in a later section of this chapter.

The ORCE and COS were among the first to consider both discrete and global measures of classroom quality, elucidating the importance of considering global classroom quality in the elementary grades as a factor in student outcomes. These elementary measures were able to capture features of elementary classrooms that were related to academic and social development of students while being content-independent (Hamre & Pianta, 2005; NICHD Early Child Care Research Network, 2002b, 2004; Pianta et al., 2007; Rimm-Kaufman et al., 2002). An analysis of results using data from the SECCYD study indicated that global ratings of the classroom using the ORCE were more related to student academic outcomes than the time-sampled teacher behaviors (Chomat-mooney et al., 2008).

This early work in observation research has moved the field toward a greater understanding of the need for measures of classroom quality to systematically examine

global measures of quality through observation. Although the majority of such tools have been developed and validated for use in early childhood general education classroom settings (Chomat-mooney et al., 2008; Maxwell et al., 2001), other observation tool development has explored the use of observation in a wider range of classroom levels (e.g., Danielson, 2011; Fish & Dane, 2000; La Paro et al., 2004; Waxman et al., 1997). The following section is an overview of the more commonly-used observational instruments designed for use in general education, followed by a review of tools designed for observation in alternative settings, (i.e., special education and intervention). The review of each observation tool will include a discussion of the purpose and content of each tool as well as a discussion of the logistical requirements for the use of each measure. Table 1 provides a summary of the review for each tool.

Table 1

Overview of Observation Tools

Observation Tool	Setting	Focus	Training Requirements	Length of Observation	Special Qualifications for Observers
Classroom Assessment Scoring System (CLASS; Pianta, et al., 2005)	General education	Emotional support, classroom organization, instructional support	16 hours	Four 30-minute cycles of observation (total: 2 hours)	None specified
Framework for Teaching (FFT; Danielson, 1997)	Primarily general education; utility in special education claimed	Planning and preparation, classroom environment, instruction (constructivist approach), professional responsibilities	12-24 hours	30 minutes-1 hour, plus time to examine pertinent artifacts	None specified, but typically designed for use by supervisors
Recognizing Effective Special Education Teachers (RESET; Johnson & Semmelroth, 2012)	Special education	Evaluating instructional practices that are evidence-based for use with students with disabilities	½ day	One lesson (15-75 minutes)	Special education teachers
Quality of Intervention Delivery and Receipt (QIDR; Harn, et al., 2011)	Small group intervention	Explicit instruction principles, instructional and behavioral management	4-6 hours	One lesson	None specified

General Education Observation Tools

In an attempt to elucidate and standardize observation methods, various observation tools have been developed and researched. Although some tools continue to focus on discrete teacher behaviors or are content-dependent, many have been developed with an emphasis on more global measures of quality that can be used in various content-independent settings (Chomat-Mooney, et al., 2008). This section will review some of the more commonly used global, content-independent tools in general education and will demonstrate the need for valid and efficient tools designed for use in intervention settings. The tools that are included in this review are among those used in the Measures of Effective Teaching Project, a large-scale teacher effectiveness study (MET; Bill and Melinda Gates Foundation, 2009). While other tools were used in the study, those included in this review are those that are considered content-independent, global measures, which facilitate greater usability in school settings.

Classroom Assessment Scoring System (CLASS). One observation tool that is commonly used in various pre-school and elementary settings is the CLASS. In fact, the CLASS has been adopted by HeadStart as its primary tool for evaluating teacher quality as part of this federal initiative (HeadStart Act, 2007). It was originally designed to assess classroom processes found to be related to student outcomes in pre-kindergarten through 3rd grade (La Paro et al., 2004; Pianta et al., 2005). The criterion and predictive validity of the CLASS have been established through multiple studies associating it with other similar tools and associating scores on the CLASS with student outcomes such as gains on standardized assessment and improved social skills (e.g., La Paro et al., 2004; Mashburn et al., 2008; Pianta et al., 2005).

Purpose. The CLASS was originally designed to be used for research purposes to understand the social-emotional climate of the classroom and how that is related to student achievement (Pianta, La Paro, et al., 2008). The instrument has been specifically used to conduct "empirically-based theories of teaching and learning that serve as the foundation for understanding education and developing new solutions" (Hamre, Pianta, Mashburn, & Downer, 2007, p. 3). Although the MET project has indicated that an ultimate purpose for the measure might be for providing feedback to improve instruction, this purpose was not within the scope of the original MET project (Joe, Tocci, Holtzman, & Williams, 2013) and has not been the focus of most other research conducted using the instrument (Hamre, Pianta, Mashburn & Downer, 2007; La Paro, Pianta, & Stuhlman, 2004; Mashburn et al., 2008; Pianta, et al., 2005).

In 2008, however, Pianta, Mashburn, Downer, Hamre, and Justice published a study using CLASS as part of a professional development and training cycle to improve classroom quality. The study was actually looking at the effectiveness of a system for professional development entitled My Teaching Partner (MTP). Within the study, two treatment conditions were utilized. The first used a system which linked videos of high-quality teacher-student interactions (based on the CLASS framework) with a consultation process using components of the CLASS as a common language for instructional quality and as a framework for providing professional development to improve instruction. The second condition provided teachers only with videos depicting high-quality instruction according to the CLASS, but did not include personal consultation. Pianta et al., (2008) reported that those teachers who received video exemplars along with personal consultation made the greater improvements in the categories of *teacher sensitivity* and

instructional learning formats than did those teachers receiving only the video exemplars. However, the authors found that the effect was greater for those teachers who were teaching classrooms in which 100% of the children were classified as experiencing poverty, and effect sizes were small to moderate across the two categories.

Content. The CLASS Framework (Hamre & Pianta, 2007) describes a theory of classroom practice derived from earlier theoretical and empirical research in educational and psychological literatures (e.g., Brophy & Good, 1986; Brophy, 1999; Gage, 1989; Pressley et al., 2003). It is framed around three broad domains of classroom interactions that are hypothesized to be important for promoting learning and social development of students: *Emotional Support, Classroom Organization,* and *Instructional Support*.

This framework is supported by previous research on classroom observation and teacher effectiveness including the work of Brophy (1999) who outlines 12 principles of effective teaching including classroom climate, opportunities to learn, curricular alignment, and student engagement. Work by Pressley and colleagues (2003), including organizing teaching strategies into creation of a motivational atmosphere, classroom management, and curricular and instructional decisions, also supports the foundational framework of the CLASS. The creators of the CLASS consider their tool to be more comprehensive than these other frameworks, however, because of a greater emphasis on social and emotional components of the classroom, specifically teacher-student interactions and relationships, as well as emphasis on instruction to enable higher-order thinking skills (Hamre, Pianta, Mashburn, & Downer, 2007).

Each of the domains *(Emotional Support, Classroom Organization, and Instructional Support)* is further subcategorized into dimensions that are explicitly

described through behavioral, observable classroom interactions between teachers and students, or among students. *Emotional Support* is broken down into four dimensions which include *Classroom Climate (Positive and Negative), Teacher Sensitivity,* and *Regard for Student Perspectives*. *Classroom Organization* includes three dimensions: *Behavior Management, Productivity,* and *Instructional Learning Formats*. *Instructional Support* includes dimensions for *Concept Development, Quality of Feedback,* and *Language Modeling*. Each of these dimensions is provided descriptors for low, middle, and high implementation of the subcategory. Ratings using the CLASS are made on a seven-point scale, ranging from "Low" to "High" on each of the ten dimensions. As an example, the "High" level of implementation of the *flexibility and student focus* subcategory has a behavioral description that indicates: "The teacher is flexible in his or her plans, goes along with students' ideas, and organizes instruction around students' interests." If a rater considered this an accurate description of the interactions being observed, a rating of "high" (six or seven) would be warranted. Conversely, an observer could rate a teacher as "low" and score a one or two if "The teacher is rigid, inflexible, and controlling in his plans and/or rarely goes along with students' ideas; most classroom activities are teacher-driven." If a rater observes that "the teacher may follow the students' lead during some periods and be more controlling during others," he or she may rate them in the "middle" category and assign a score of three, four, or five. This same approach is used across the other dimensions as well. For the complete rating scale descriptors on the CLASS, see Pianta, La Paro, and Hamre (2008).

Training. To ensure that raters can provide reliable and accurate scores using the CLASS, developers indicate that sixteen hours of training is required. Developers of the

CLASS indicate that observers should have some teaching experience, however, it has been found that teachers and administrators with the most experience are often less reliable due to preconceived notions regarding effective teaching that may not align with elements of the CLASS (Hamre, Goffin, & Kraft-Sayre, 2009). A manual containing descriptions of each of the domains and dimensions, to be read prior to training, is provided to trainees. The two-day workshop consists of guided practice with videotaped classroom footage and an extensive videotaped reliability test, involving either five or six cycles of 20-44 minute observations. With this level of training, an average interrater reliability (within one point of master coders) of .87 has been reported (Pianta, Mashburn, et al., 2008).

Observation duration. Developers of the CLASS recommend a minimum of two hours of observation, in the form of four, 30-minute cycles, in order to obtain a reliable measure of classroom quality (Chomat-mooney et al., 2008). In general, it is also recommended that multiple observation cycles of each classroom, across different points in the school year, be obtained in order to confidently determine a level of quality within the classroom (Kane & Staiger, 2012).

Framework for Teaching (FFT). The Framework for Teaching (FFT; Danielson, 1997) is another observation tool that is also widely used within general education settings. Twenty states and various school districts in the United States have adopted the FFT as a means for evaluating teachers (Hansen, Lemke, & Sorensen, 2013). Numerous studies have indicated predictive validity of the measure on student learning outcomes (Borman, Kimball, Borman, & Kimball, 2005; Heneman, Milanowski, Kimball, & Odden, 2006; Holtzapple, 2003; Kane, Taylor, Tyler, & Wooten, 2010;

Milanowski, 2004). For example, Kane, Taylor, Tyler, and Wooten (2010) found that a one point increase in FFT scores accounted for achievement gains of one-fifth and one-sixth of a standard deviation for reading and math, respectively. In another example, Heneman, et al., (2006) used correlational research to determine the relationship between teacher performance on the FFT and student achievement on both reading and math across four sites. In the area of reading achievement, scores on the FFT correlated with reading achievement with an average correlation of .29, with a range of correlations from .22 to .37. Correlations of FFT scores with mathematics achievement averaged .23 with a range of correlations from .11 to .32 across the four sites.

Purpose. The Framework for Teaching (FFT) was first published by the Association for Supervision and Curriculum Development in 1996. It was an extension of the *Praxis III: Classroom Performance Assessments* that had been developed over a period of six year (1987-1993) by Educational Testing Services (ETS) as an observation-based method to evaluate quality of pre-service teachers for the purpose of licensure. The FFT expanded on ETS' work by including skills of teaching required by all teachers, not just pre-service teachers (Danielson, 2011). Danielson (2007) maintains that an evaluation system must serve two purposes, to: a) ensure teacher quality and b) inform professional development. The FFT was designed to reflect current notions of "best practices" and to function as both a formative and summative evaluation tool (Danielson & McGreal, 2000).

Content. The FFT is considered by developers to be a contemporary form of observation that focuses on constructive approaches to teaching (Danielson, 1996). The framework is based upon an underlying notion that teachers honor and nurture the

students' natural impulse to construct new understandings (Brooks & Brooks, 1999). The knowledge base for the original ETS version of the framework for teaching was developed around three information sources: wisdom of experienced teachers, theory and data of educational researchers, and requirements for licensure from various states (Danielson, 2007). Surveys were used to access information from experienced teachers to perform job analyses of teachers from elementary, middle, and high school. Extensive literature searches were used to review and synthesize research on effective teaching and requirements of state licensing agencies were analyzed and incorporated within the ETS version of what would later become the FFT. In accordance with state licensing agencies, the developers designed the FFT to be aligned with the Interstate New Teachers Assessment and Support Consortium (InTASC; Council of Chief State School Officers, 2011), a set of standards used to measure competency of pre-service teachers in many teacher preparation programs throughout the United States. The latest edition of the FFT has also been modified in an effort to reflect the instructional implications of the Common Core State Standards (CCSS; Danielson, 2013).

The FFT is organized around four broad domains: *Planning and Preparation, Classroom Environment, Instruction,* and *Professional Responsibilities*. Each of these domains consists of five or six components. These components are further defined through elements related to each component. For instance, within the *Planning and Preparation* domain, there are six components: *demonstrating knowledge of content and pedagogy, demonstrating knowledge of students, setting instructional outcomes, demonstrating knowledge of resources, designing coherent instruction, and designing student assessments*. Each of these components is further defined with additional

elements. The scoring rubric contains four possible levels of implementation: Level 1, Unsatisfactory; Level 2, Basic; Level 3, Proficient; and Level 4, Distinguished. Within this rubric, specific examples and detailed explanations are provided to aid in assigning scores during observation.

Research informing the first domain (*Planning and Preparation*) was derived from multiple sources and highlights organizational skills, planning, content and pedagogical knowledge, using students' prior knowledge, having high expectations, and establishing clear goals (Brooks & Brooks, 1999; Jackson & Davis, 2000; Marzano, 2004; Schmoker, 1999; Shulman, 1987; Sykes & Bird, 1992; Wiggins & McTighe, 1998). The second domain, *Classroom Environment*, draws upon research indicating that teachers must master at least basic levels of classroom management (i.e., creating routines and procedures, building an efficient and functional physical environment, and establishing norms and expectations for student behavior) prior to becoming skilled at providing instruction (Evertson & Harris, 1992; Jackson & Davis, 2000; Jensen, 1998; Tomlinson, 1999). *Instruction*, the third domain of the FFT is designed to reflect the emphasis on teaching for understanding and conceptual learning and is based on the premise that children benefit most when allowed to "construct" new learning based on prior knowledge (Danielson, 2007). This domain was informed by research highlighting the importance of communicating expectations and goals, a need for flexibility, questioning and discussion skills, and assessment practices (Brooks & Brooks, 1999; Skowron, 2001; Tomlinson, 1999). The final domain, *Professional Responsibilities*, is an attempt to measure the full range of responsibilities that constitute teaching, including commitment to student learning, systematic reflection of teaching practice, collaboration

in a learning community, and effective parent involvement (Colton & Sparks-Langer, 1992; Danielson, 2007; Jackson & Davis, 2000; Ross & Regan, 1993; Stronge, 2005).

Training. According to McClellan, Atkinson, and Danielson (2012), training should include a minimum of 3-4 hours of an introduction to the tool, including the process for observation and an overview of the tool, as well as training to overcome potential bias. An in-depth training of the content of the tool requires between 12 and 24 hours. Embedded within this training is an additional 12 hours for practice scoring of clips for each of the domains. Lastly, observers should spend between eight and ten hours scoring full-length practice videos. Overall, the training should be between forty and fifty hours in order to ensure reliability. The authors indicate that inter-rater reliability should be at a level of .80 or higher following training. The authors do not offer suggestions on levels of experience preferred for observers.

Observation duration. The FFT was designed to be used for full-length observations of lessons, ranging from 30 minutes to one hour. However, some of the components (e.g., planning and preparation, and professional responsibilities) require additional time to examine artifacts such as lesson plans, inspect evidence of participation in professional development opportunities, and investigate the nature of interactions with colleagues (Danielson, 2007).

Special Education and Intervention Observation Tools

Measuring teacher effectiveness within the context of special education and other intervention settings can be quite complex (Brownell et al., 2009). Since the goal of special education/intervention instruction is to provide more targeted and/or individualized instruction, tools designed for use in general education settings may be

inappropriate (Johnson & Semmelroth, 2012; Jones & Brownell, 2013). The FFT claims to have utility within the context of special education, acknowledging that there might be slight variations in the delivery and responsibilities of specialists, but that, "fundamentally, they are all teachers of students" (Danielson, 2007; p. 109), making the framework applicable to a variety of settings. However, as Jones and Brownell (2014) explain, instruction in a special education or intervention environment must be designed to focus on skills that are likely very difficult for the student to grasp requiring teacher-directed, intensive, and repetitive tasks for students to acquire the knowledge and skills being taught. This teacher-directed approach is in direct contrast to the more constructivist framework that the CLASS and FFT tools advocate and measure.

Because of this difference of definition of effective instruction, some researchers have sought to develop and validate tools measuring the types of instruction that are more likely seen in intervention settings (Harn, Forbes-Spear, Fritz, & Berg, 2011; Johnson & Semmelroth, 2013). This section will outline two tools specifically developed to measure instruction within the special education or intervention context, the Recognizing Effective Special Education Teachers Observation Tool (RESET; Johnson & Semmelroth, 2013) and the Quality of Intervention Delivery and Receipt (QIDR; Harn et al., 2011).

Recognizing Effective Special Education Teachers Observation Tool (RESET). The RESET Observation Tool (Johnson & Semmelroth, 2012) was specially designed to measure effectiveness of special education teachers and take into account the more varied settings and instructional strategies used by special education teachers. The developers set out to design an observation tool that was a systematic observation

approach, aligned with evidence-based practices for students with disabilities, and that could serve as an alternative to the FFT, which the authors contend may not be aligned with the research base around best practices for students with disabilities, and may endorse practices that do not lead to improved outcomes for students with disabilities (Johnson & Semmelroth, 2012; Semmelroth & Johnson, 2013).

Purpose. Following the lead of Danielson (2007), the developers of the RESET observation tool sought to develop a tool that could provide feedback that could serve the same purposes as the FFT (i.e., to ensure teacher quality and promote professional development), but specifically in the special education context. The developers aimed to develop a tool that addressed the diversity found within special education classrooms and acknowledged the unique struggles found in the special education profession (Semmelroth, 2013). The RESET system was also designed to provide feedback on specific instructional practices to allow special education teachers to improve their practice (Semmelroth, 2013).

Content. The content of the RESET observation tool is based on Danielson's (2007) framework with a focus on Domain 3, *Instruction*. It differs from the FFT, however, in that it includes more explicit criteria for evaluating evidence-based instructional practices appropriate for students with disabilities (Semmelroth, 2013). The tool was developed within a framework that defines special education teachers as those who are able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth (Johnson & Semmelroth, 2012).

The RESET observation tool was developed through an extensive review of research within special education. Three sources informed the content of the tool: a) Danielson's FFT (2007), Domain 3: *Instruction*; b) Council for Exceptional Children (CEC) professional Standards for Special Education Teachers (2009); and c) a meta-review of literature on effective special education instructional practice (Semmelroth, 2013). Through this review of research, the developers created a tool designed to be flexible enough to be used across various special education settings (e.g., inclusive settings, small-group direct instruction, team-teaching) and addressing the needs of students with various disabilities (Semmelroth & Johnson, 2013).

The initial version of the RESET consists of three main parts: Lesson Overview (introduction), Lesson Components (instructional practices), and Lesson Summary (conclusion). Three different evidence-based instructional practices are included in the RESET tool*: direct, explicit instruction, whole-group instruction*, and *discrete trial teaching*. There are between 28 and 67 items on the RESET depending on the number of instructional practices being observed. The tool is web-based, operating on a direct logic system (i.e., some questions only appear if previous questions have been answered in a pre-defined way; Johnson & Semmelroth, 2012). For instance, if the observer indicates that the lesson being observed is employing direct instruction, only scoring related to direct instruction is revealed to the observer (Johnson & Semmelroth, 2012). In that instance, observers would be using the *Explicit, Direct Instruction* component of Subscale 2: *EBP Implementation*. Within the *Explicit, Direct Instruction* component, more specific sub-headings are evaluated: *a) Organized Instruction; b) Sequenced Instruction; c) Student Participation; d) Scaffolding;* and e*) Assessment*. The second

evidence-based instructional practice included in the RESET tool is the "*Whole Group Instruction*" component which includes subheadings of *a) Individualized Instruction, b) Skill Development, c) Student Engagement,* and *d) Feedback and Assessment*. The third evidence-based instructional practice included in the RESET tool is the *Discrete Trial Teaching* component including subheadings of *Antecedent, Response, Consequence,* and *Intertrial Interval* (ITI).

The rubric for scoring of the RESET is based on Danielson's (2007) four-point scale: zero (unsatisfactory), one (basic), two (proficient), three (distinguished). Within the rubric, developers have included behavioral descriptors to aid observers in assigning a score. For example, within the *Whole Group Instruction* component: *Student Engagement*, two descriptors for the score of zero are provided: "The teacher provides little to no opportunities for guided and independent practice for students," and "The teacher provides little to no opportunities for students to participate in classroom activities." Conversely, a score of three on this component indicates "The teacher provides for individualized opportunities for guided and independent student practice for all students," and "The teacher has created a learning environment that encourages active participation from all students, as well as maintains active levels of self-determination and self-advocacy." For more excerpts from the RESET observation tool, see Semmelroth (2013).

Training. For training, observers are provided a manual outlining the components of the RESET observation tool. A half-day training presentation is provided to orient observers to the tool and provide opportunities for explaining the manual and the observation tool (Semmelroth & Johnson, 2013). Following the presentation, observers

are given the opportunity to view a practice video as a group activity. Observers rate the video and differences in scores are discussed until consensus is reached (Johnson & Semmelroth, 2012). Following this, observers rate a second video and scores are reviewed in a whole group activity. Developers reported an interrater agreement of .72 to .95 during training, measured both as a holistic score and by each subscale (Semmelroth & Johnson, 2013). The developers of the RESET tool sought only special education teachers for the initial training during this pilot version of the RESET. The teachers ranged in experience from five to thirty years with an average of twelve years of teaching experience.

Observation duration. Developers designed the RESET observation tool to be used with video recordings of single lessons. The mean time of each video used during development of the tool was 25 minutes, with videos ranging from 17 to 72 minutes. Regardless of video length, the videos were representative of one lesson and are to be observed and rated in their entirety (Semmelroth & Johnson, 2013).

Quality of Intervention Delivery and Receipt (QIDR). Similar to the RESET observation tool, the QIDR is also designed to be used in settings other than general education classrooms (Harn, et al., 2011). Unlike the RESET, however, the QIDR is designed to measure only small group, direct, explicit instruction that is typically found in intervention settings. It was not developed specifically for use in special education, but for all intervention settings which involve small group instruction, independent of content area (i.e., reading, math).

Purpose. The QIDR tool (Harn, Forbes-Spear, Fritz & Berg, 2012) was developed for two main purposes. The first was to measure the quality of small group intervention

delivery to identify and measure specific elements of instruction that are related to outcomes and accelerate student learning. For each of the 15 specific instructional skills measured on the QIDR, a rubric was created to assess the quality of how that instructional skill was delivered on a scale of 0-3. By measuring targeted instructional behaviors with a qualitative lens and in a systematic manner, specific instructional behaviors could be examined to identify potential research areas to focus on to better support students. The second, more applied, purpose for developing the QIDR was to provide a tool for principals and coaches to use to provide specific feedback to interventionists to drive instructional improvement. Although the tool had dual purposes, each purpose required that the tool measure multiple facets of instructional delivery and student behavior.

Content. To meet these purposes, developers looked to various sources to determine what aspects of instruction were most related to improved student outcomes. The content of the QIDR observation tool is not dependent on specific academic instructional content (i.e., reading, math), but instead is based on instructional principles that have evidence of increasing student achievement. In an intervention setting, instructional behaviors related to systematic, explicit instruction have shown positive effective sizes (Gersten et al., 1997; Swanson, 1999), indicating that instruction in these settings must be explicit, intensive and supportive (Torgesen, 2002). Therefore, items within the QIDR are derived from instructional principles commonly used in intervention settings to accelerate academic achievement in students who are at-risk or in need of remediation.

The main elements of the QIDR were developed to reflect key instructional elements necessary for providing explicit instruction (Archer & Hughes, 2011; Brophy & Good, 1986; Brophy, 1986; Rosenshine & Stevens, 1986). Similar to how the CLASS was developed from the effective teacher research completed during the 1970s and 1980s, specific instructional behaviors were identified as common among teachers consistently getting positive outcomes from their students (Brophy & Good, 1986; Brophy, 1986; Dunkin & Biddle, 1974; Medley, 1979; Rosenshine & Stevens, 1986; Rosenshine, 1971). Through systematic classroom observations, common instructional behaviors were found to correlate with student outcomes. For example, Brophy and Good (1986), through a synthesis of previous research, found that the most consistently replicated finding in observation research linked student achievement to opportunity to learn material and the degree to which teachers provided that opportunity through active participation and moving students through curriculum at a brisk pace. Achieving these opportunities was found to be related to high teacher expectations and classroom management that provided organized environments that maximized student engaged time (Brophy & Good, 1986).

Items on the QIDR related to the work of Brophy and Good (1986) include providing specific feedback for correct and incorrect responses, using clear and consistent wording, and modulation of lesson pacing. Further items were informed by the work of Rosenshine and Stevens (1986) including providing multiple and varied opportunities for guided and independent practice, providing frequent modeling, and ensuring students achieve mastery before moving on to new concepts. Archer and Hughes (2011) indicate that there are 16 elements of explicit instruction that are important for ensuring positive

student outcomes. Many of these elements are informed by the previously-mentioned work, but some items in the QIDR reflect Archer and Hughes' (2011) extension of these instructional principles, including organizing instruction systematically, and declaring academic and behavioral expectations.

In addition to research-based instructional elements, the development team included elements on the QIDR related to student and group management of behavior. Research indicates that organization and management, along with positive social-emotional climate, may help increase engagement and opportunities to learn, thus positively impacting student academic outcomes (Brophy, 1986; Cameron et al., 2005; Connor et al., 2009; Hamre & Pianta, 2005). Based on these findings, items related to organization and emotional support, including organization of materials, familiarity and preparation of lessons, smooth transitions, and teacher's responsiveness to the emotional needs of children are also measured.

The wording of items on the QIDR was also carefully developed for two purposes: 1) to systematically and reliably measure across observers and 2) to give precise feedback to interventionists that could guide instructional improvement. Each instructional and behavioral element is behaviorally operationalized across the four levels of rating. For example, one of the items on the QIDR related to management is "Teacher appropriately responds to problem behavior." The item is described in detail with examples such as "including off task behaviors; emphasizes success while providing descriptive, corrective feedback; positively reinforces to get students back on track." In other words, coaches and interventionists could use descriptions on the scoring rubric to

determine specific actions to improve a score on a certain item, thus potentially improving instructional quality over time.

For each of the items on the QIDR, observers provide scores using a Likert scale, ranging from zero, "Not implemented," to 3, "Expert implementation." Each of the levels of implementation is behaviorally operationalized with use of examples and frequency (when appropriate) to distinguish one level from the others. For instance, for the item related to responding appropriately to problem behavior, an interventionist would receive a score of zero to three based on the observers perception related to the following: 0= "Teacher does not appropriately respond to problem behavior across multiple students. Teacher primarily provides negative feedback or ignores problem behavior for extended period of time (resulting in limited student participation, e.g., more than 20% of activity); 1= "Teacher sometimes appropriately responds to problem behavior. Teacher provides some positive or corrective feedback but does not regularly emphasize success. Teacher may have difficulty consistently responding to one student's problem behavior but sometimes responds appropriately to other students"; 2= "Teacher typically responds appropriately to problem behavior by emphasizing success and providing neutral corrective feedback for most students. ***Or*** no problem behavior occurs during the instruction"; and 3= "Teacher consistently responds appropriately to problem behavior by emphasizing success and providing descriptive corrective feedback as needed for all students. For example, teacher "catches" students engaging in appropriate behavior and provides descriptive positive feedback to encourage appropriate behavior." All items and their operationalized definitions are included in the rubric (see Appendix). Preliminary

evidence on the QIDR indicates that it is significantly related and predictive of student outcomes (Forbes-Spear, 2014).

Training. The initial training on the use of the QIDR requires four to six hours. The training consists of an overview of the observation tool, explanation and examples of each element, guided practice of scoring on each element using segments of videos of small group instruction, as well as feedback on scoring accuracy for each element. All training videos were originally scored by the original QIDR team who independently coded each training video, discussed any disagreements, and used a consensus-building approach to come to agreement on "true" scores. Raters are then provided the opportunity to practice scoring using all elements of the QIDR while watching a recorded 30-minute intervention session. This guided practice is followed by immediate feedback on scoring accuracy across all elements. Raters who provide scores that are no more than one point off true scores (adjacent scores) according to the Likert scale for each item, are considered on-track to obtain acceptable reliability, and given the opportunity to independently score three check-out videos. If raters do not provide adjacent scores on all items, re-training, including discussion, re-visiting key elements of the scoring rubric, and additional guided practice, is provided.

After raters score each check-out video, their scores are compared to true scores. Raters who obtained an intraclass correlation (*ICC*) of .6 or higher compared to the derived true scores (correlation of .7 or higher between rater and true score) after each video were cleared to independently score the next check-out video. Those who fell below this cut-off were provided re-training followed by additional check-out videos until an acceptable *ICC* was reached. No coder needed more than two additional checks

to obtain reliability. Once an observer demonstrated consistency and agreement in scoring with the true-scored videos, they could score videos independently. The training can be delivered in a face-to-face setting, which allows for discussion and support as needed, or through an online training module which allows raters to train at their own pace followed by telephone or email support as needed for retraining.

Raters using the QIDR during the pilot phase of the development of the tool had various backgrounds and levels of experience within the field of education. Some observers were undergraduate and graduate students with little or no teaching experience, while others were teachers with multiple years of teaching experience. Regardless of level of experience, all raters were able to be successfully trained to use the QIDR reliably which indicates that the range of observers could potentially be quite diverse and still obtain reliable measures of instructional quality. Despite the range of backgrounds within pilot training, need for retraining was relatively limited. Only one of three coders required informal retraining through discussion to gain reliability. Formal retraining was not required for any team members.

Observation Duration. The QIDR was initially designed and used to measure instructional quality across an entire intervention lesson. Given the intensive nature of intervention instruction, these sessions typically run between 15 and 30 minutes.

Maximizing Time for Observation and Feedback

Each of the observation tools discussed in the previous section has the potential to provide important insight into instructional quality. A major challenge in using many of these tools is the extensive time necessary to train raters and complete observations (Gargani & Strong, 2014). The extended time required to carry out these tasks may make

the use of some of these tools prohibitive for the purpose of providing the regular feedback necessary to improve instruction. Two groups of researchers have begun to investigate alternatives to the current observation systems that can maximize efficiency of observation while balancing the validity of the data (Gargani & Strong, 2014; Pratt & Logan, 2014).

Proximal and distal measures of quality using short observations. Pratt and Logan (2014) conducted a study to investigate the effects of *Let's Know* (Language and Reading Research Consortium, in press), a supplemental curriculum for pre-kindergarten through 3rd grade, on teachers' use of language-related comprehension supports. Researchers set out to examine two questions: 1) how the *Let's Know* curriculum impacted teachers' use of language-related comprehension supports; and 2) how the *Let's Know* curriculum impacted quality of general teacher instructional delivery.

To examine the first question, researchers developed an observation tool, *Snippets*, as a proximal measure of teachers' use of language-focused comprehension supports. The tool was designed to observe very short samples, or "snippets", of instruction (i.e., six-minute segments). *Snippets* contains eighteen items related to reading comprehension skills known to be important for pre-kindergarten through third grade students (e.g., prediction, inference, summarizing, main idea).

Within the study, two six-minute segments of video, recorded during a 90-minute reading block, were coded using the *Snippets* tool. One of the six-minute segments was extracted from a 30-minute lesson in which the teacher was delivering the supplemental *Let's Know* (Language and Reading Research Consortium, in press) curriculum. The other six-minute segment was taken from another part of the remaining hour of the same

reading block to determine whether or not these language-based comprehension supports were being used during normal language-arts instruction, but outside of the *Let's Know* lessons. Each six-minute video segment was coded using an interval-based scheme in which observers coded 12, 30-second intervals for the presence or absence of any of the 18 language-focused comprehension supports. Therefore, in one six-minute segment, scores for each support could range from 0-12. Although the authors do not indicate how much training was required to gain reliability, reliability was assessed for 14% of the segments coded and obtained overall exact agreement of 98%, which translated into a Kappa calculation of .86.

To measure quality of the instruction being delivered, Pratt and Logan (2014) also used the CLASS (Pianta, Mashburn, et al., 2008) as a global measure of quality of implementation, using time segments that were half those specified by CLASS protocol. Observers rated four 15-minute segments of the 90-minute reading block (two during the *Let's Know* lesson, two outside of the *Let's Know* lesson) rather than a minimum four cycles of 30 minutes each as indicated by the CLASS protocol. The instructional support (IS) domain of the CLASS was used to code the 15-minute segments to determine if the *Let's Know* curriculum also impacted ratings using the instructional support domain of the CLASS. Even though observers were only coding a portion of the reading block, observers will able to obtain 89% agreement based on the percentage of within-one agreement (the predominant approach used to assess reliability with the CLASS). During *Let's Know* lessons, teachers scored significantly higher in instructional support than they did in segments scored outside *Let's Know* lessons. This difference demonstrated evidence for discriminate validity indicating that the differentiated instruction and use of

language-based comprehension supports present during the *Let's Know* intervention provided increased quality of instructional supports, as measured by the CLASS, that were not present outside of the intervention.

Maximizing the efficiency of observation. Gargani and Strong (2014) believed that an observation tool needed to be developed that could measure teacher effectiveness quickly and efficiently to expedite teacher evaluation systems. These researchers felt that popular observation tools such as the CLASS (Pianta et al., 2008) and the FFT (Danielson, 1997), while comprehensive and reflective of a broad set of standards for good teaching, were not necessarily designed with the observer in mind. They felt that the training and use of these tools might be too cumbersome for practical use within schools. To remedy this situation, they set out to design a more efficient observation tool that could predict the ability of teachers to raise the achievement of their students as a part of a system of teacher evaluation (Gargani &Strong, 2014).

The result of their efforts was an observation tool called the Rapid Assessment of Teacher Effectiveness (RATE; Gargani & Strong, 2014; Strong, 2011). The current version of the RATE contains six rubric items. The authors do not contend that the six items within the scoring rubric for RATE define good teaching. Their main goal in the development of RATE was to determine if six deliberately-chosen items were sufficient to predict student learning on standardized tests (Gargani & Strong, 2014).

The RATE observation tool contains six items for evaluating instruction. The items were derived, in part, from items on the CLASS (Pianta et al., 2008) as well as items derived from previous work in which raters were asked to classify teachers according to whether the raters felt that students achieved above or below average

depending on the teacher being observed (Strong, Gargani, & Hacifazlioglu, 2011). As a part of this early work, raters were polled to determine what factors most influenced their judgements. The raters cited student engagement, teaching strategies, and math knowledge as the most important factors, with student engagement being the most frequently cited. The results of this polling informed additional items within the rubric. The items within the rubric relate to lesson objectives, instructional delivery, questioning strategies, clarity of presentation, time on task, and level of student understanding. Each of the items is scored on a scale of one to three, with behavioral descriptions for each level of implementation. To provide scores for teacher quality, raters viewed only the first 20 minutes of a videotaped lesson and were allowed ten additional minutes to create scores using the rubric.

One of the purposes of the development of RATE was to provide a tool that requires minimal training while still producing reliable scores that are predictive of student outcomes (Gargani & Strong, 2014). For this reason, training is limited to one 2-hour training session. Throughout the series of validation studies, researchers purposely chose observers with widely varied backgrounds. Some were undergraduate and graduate research assistants with no teaching experience, while others were teachers with experience using other rating systems. The validation studies were designed to determine if the tool provided scores that were predictive of increases in student learning as assessed using value-added measures (VAMs) and if it could be used reliably. The researchers reported that across five of the studies, observers were able to classify teachers as either high or low performing, according to their VAMs, between 70 and 78 percent of the time. Across five separate studies, interrater reliability was obtained using

intraclass correlations (*ICC*s) in the same way as the MET study. The range of *ICC*s for independent scoring was .31 to .92 with an average of .65 across the five studies. This average places IRR for this tool in the good category (between .60 and .74) according to commonly-cited cutoffs for qualitative ratings of agreement (Cichetti, 1994).

Although these studies have varying purposes from each other and the current study, the work of these researchers can help to inform the next series of research studies to examine methods for how to more efficiently and validly assess instructional quality that is related to student outcomes. The current study utilized the QIDR (Harn et al., 2011) and systematically observed short (i.e., 10 minutes) segments of instruction (i.e., beginning, middle, end) to determine how these segments relate to the overall intervention time as well as to student outcomes.

Summary and Conclusions

The current climate in education has placed greater emphasis on evaluating the effectiveness of teachers to ensure students are receiving the highest quality instruction (McGuinn, 2012; National Association of Teacher Quality, 2012). Evaluating teacher effectiveness is viewed as a means for ensuring teacher quality and determining needs for professional development (Danielson, 2007). While multiple methods for evaluation have been researched and used in schools, observation remains one of the most direct ways to evaluate the quality of classrooms and instruction (Chomat-mooney et al., 2008). Multiple tools have been developed and tested, primarily with a focus on general education, but these tools are often complex and cumbersome, creating many challenges in implementation, particularly in using these tools in intervention settings (Gargani & Strong, 2014).

One major challenge with using general education observation tools within intervention settings is the focus on measurement of instructional strategies that may not be appropriate or effective in settings in which a different type of instruction is expected, valued, and needed (i.e., direct, explicit instruction; Johnson & Semmelroth, 2012; Jones & Brownell, 2014). If these general education tools are used in intervention settings, interventionists may be misidentified as being in need of professional development when in fact the type of instruction they are providing is simply not a match for the observation tool, yet effective at improving student outcomes. Another major difficulty with current observation tools is the resource-intensive nature of both training and observation (Gargani & Strong, 2014). Due to the large scope of instruction that these tools are attempting to measure, they require extensive, time-consuming training to ensure reliability of raters. In addition, observations themselves can be time-consuming leaving little time for providing necessary feedback to improve instruction (The New Teacher Project, 2013). In the current state of education, resources within schools are already stretched thin so requiring administrators and coaches to complete extensive training (e.g., approximately 16 hours) and spend considerable time carrying out observations (e.g., four 30-minute observation cycles) and providing feedback to improve instruction is prohibitive. Students who are receiving intervention are arguably in need of the best quality instruction, but because of these challenges, interventionists are unlikely to be provided feedback that is frequent enough to improve instruction.

To optimize the utility of observation tools to ensure greater student outcomes in intervention settings, they must be specific enough to provide feedback that can improve instruction, yet efficient enough to allow for succinct training and relatively short

observation periods. Currently, there is no observation tool that can efficiently provide specific feedback that can be used in a responsive instructional cycle. The development of the QIDR (Harn, et al., 2012) is an important step in providing a tool that can provide specific feedback to improve intervention instruction. The purpose of the current study is to determine if the QIDR tool can be used to garner reliable and valid measures of instructional quality without having to watch an entire instructional session (e.g., 25-33% of intervention).

The next chapter will describe the methods as well as the population and data set that were used to explore these questions.

CHAPTER III

RESEARCH METHODS

Implementation science has revealed that one of the factors that can improve implementation of intervention is providing frequent feedback to interventionists (Fixsen et al., 2013; Odom, 2008). Providing feedback is likely to improve intervention instruction that can, in turn, ensure that students in most need are receiving the highest quality instruction (Connor, 2013). To provide this frequent feedback, it will be necessary to make the process more efficient while still maintaining a high level of quality of such feedback. In an earlier study, the QIDR observation tool was used to measure instructional quality of entire lessons (i.e., 20-30 minutes) delivered by seven interventionists to 35 kindergarten students considered at-risk for reading difficulties. Sixty-four videotaped full-length lessons were used in the earlier study and coded using the QIDR. The current study utilized the same instructional videos and the QIDR to measure quality of delivery from systematically sampled lesson segments (i.e., beginning, middle, and end) of the same videos. These samples were then compared to the full-length lessons previously-coded. This study employed Classical Test Theory (which states that all observed scores are comprised of a true score and error, both random and systematic) to determine whether an observation tool designed to measure intervention implementation [Quality of Intervention Delivery and Receipt (QIDR)] could be used reliably on short samples (i.e., ten minutes) of a lesson. This chapter provides an overview of the existing data set, as well as the methods that were used to analyze the data to answer the following questions:

1) Can adequate inter-rater reliability (IRR) be obtained after observing ten minutes of full-length intervention lessons?
2) Using the QIDR, what is the relationship between scores obtained watching the full lesson versus sampling ten minutes of the lesson?
3) To what extent does the relationship between QIDR ratings obtained watching the full lesson, versus sampling ten minutes of the lesson, depend on lesson segment (i.e., beginning, middle, or end) or on intervention phase (i.e., 2nd week, 5th week, or 8th week)? In other words, are correlations between the ratings systematically stronger or weaker based on lesson segment or intervention phase?
4) Which QIDR ratings (full lesson vs. ten-minute sample; beginning, middle, end; intervention phase) account for the most variance in student outcomes?

Setting and Participants

Setting. The original study, from which the videos for the current study were obtained, used data collected from two elementary schools in a school district in a mid-size city in the Pacific Northwest. The first school involved in the study had 646 kindergarten through 8th grade students in the 2011-2012 school year. According to the state department of education school report card for 2011-2012, 58% were considered economically disadvantaged, 17% of students were classified as students with disabilities, and 11% were considered limited English proficient, participating in English as a second language programs. The demographic make-up of this school included 69% White (not of Hispanic origin), 19% Hispanic origin, 4% Asian/Pacific Islander, 3% American Indian/Alaskan Native, 2% Black (not of Hispanic origin), and 3% multi-racial/multi-ethnic. The second school had an enrollment of 285 students in kindergarten through fifth

grade during the 2011-2012 school year. According to the state department of education report card, 79% percent of these students were considered economically disadvantaged, 15% were students with disabilities, and 15% were limited English proficient, with 13% of students participating in English as a second language programs. The demographic make-up included 63% White (not of Hispanic origin), 24% Hispanic origin, 4% Black (not of Hispanic origin), 3% Asian/Pacific Islander, and 5% multi-racial/multi-ethnic.

Student participants. Kindergarten children in the two schools received half-day kindergarten and those who were identified as at-risk for reading difficulties were given the opportunity to participate in an intervention program that was entitled *Super K*. Students in the *Super K* program either stayed after or arrived early for their regular classroom day to receive approximately 30 minutes of reading intervention.

Students were selected for the *Super K* program through use of Dynamic Indicators of Basic Early Literacy Skills (*DIBELS;* Good et al., 2002). DIBELS Letter Naming Fluency (LNF) and Initial Sounds Fluency (ISF) were administered to all kindergarten students in the fall to identify those in need of intervention. LNF presents students with a page of randomly-arranged upper and lowercase letters and asks students to name as many as possible in one minute. One-month, alternate-form reliability of LNF is .88 in kindergarten (Good et al., 2004). Criterion-related validity with the Woodcock-Johnson Psycho-Educational Battery-Revised Readiness Cluster standard score is .70 in kindergarten (Good, et al., 2004) and predictive validity of kindergarten LNF with first grade Woodcock-Johnson Psycho-Educational Battery-Revised Readiness Cluster standard score is .65 (Good, et al., 2004).

Those students who scored less than three sounds or letters per minute on ISF and LNF, respectively, were considered most at-risk and invited to participate in the *Super K* program. As a result of the screening process, 37 students across the two schools were included in the *Super K* program. Consent was obtained from 35 students, but four students moved prior to the end of the intervention, leaving a remaining final sample of 31 students. Seven of the 31 students received special education services and eleven were considered English language learners (ELLs).

Interventionists. In the original study, seven instructional assistants delivered instruction during the *Super K* intervention program. In these schools, it was typical for instructional assistants to deliver intervention programs under the guidance of a certified teacher. The interventionists involved in this study had between nine and 15 years of experience as instructional assistants and three to 14 years of experience using the specific reading programs delivered during the *Super K* program.

Observers. During the original study involving the *Super K* program, a team of seven observers coded all 64 videos within the data set in their full-length form using the QIDR. For the current study, an additional team of observers were trained to use the QIDR to code ten-minute samples of the same videos for comparison. This team was originally comprised of seven observers, with two from the original observation team, but reliability issues reduced the final number of coders to five, maintaining the two experienced coders. Given that observers in the original study had a wide variety of backgrounds and levels of experience in the field of education, observers were recruited for this study from both graduate students and practicing teachers. The eliminated coders included one graduate student and one practicing general education teacher. The final set

of coders included three graduate students with former elementary-level teaching experience, one graduate student with no teaching experience, and one practicing general education elementary teacher.

Intervention Programs

***Super K* intervention.** Students in the *Super K* program received instruction using either Early Reading Intervention (ERI; Simmons & Kame'enui, 2003), Reading Mastery (Engelmann et al., 2002), or a combination of both programs. Both of these programs are scripted and use explicit instruction principles (Archer & Hughes, 2011; Brophy, 1988; Brophy & Good, 1986; Rosenshine & Stevens, 1986) focusing on development of phonological awareness and alphabetic principle skills. The intervention was supplemental to the school's core reading program, and occurred either before or after students' regular instructional day. Intervention instruction was delivered in small groups of 3-5 students, with an average of 34 intervention sessions (range 28-41) provided to participants.

Measures

Instructional implementation measure. The QIDR (Harn et al., 2012) was used to measure instructional implementation across the videos. The QIDR is designed to be used to evaluate overall quality of an intervention based on key elements of explicit instruction. The QIDR is an observation instrument with a global approach to measuring instructional quality that is looking at multiple facets of instruction within the context of small group intervention. Items within the QIDR reflect key instructional elements of explicit instruction which are commonly used in intervention settings to accelerate academic achievement in students who are at-risk or in need of remediation (e.g., Gersten

et al., 1997; Swanson, 1999). See Chapter 2 for a complete description of the QIDR observation tool, as well as preliminary validation information.

Student outcome measure. The student outcome measure that was used to investigate the fourth research question for this project was the Word Attack (WAT) subtest of the Woodcock Reading Mastery Tests—Revised (WMRT-R; Woodcock, 1987). Scores were obtained at the beginning and end of the intervention, but for purposes of this investigation, only post-test scores were used. The WAT subtest assesses phonetic decoding skills by presenting real and nonsense words of increasing difficulty for students to read aloud. The publisher reports a split-half reliability range of 0.86-0.94 for the WAT subtest. Concurrent validity ranges for total reading of the WRMT-R are reported to be from 0.85-0.91 when compared to the full scale reading test of the Woodcock-Johnson Psycho-Educational Battery for grades 1, 3, and 5 (Woodcock & Johnson, 1989).

Video Data Set

Full-length videos. During the initial study, intervention sessions for a total of eight groups were recorded once each week, with one interventionist delivering instruction to two different groups. Each group recorded between seven and nine sessions resulting in a video data set of 64 videos. The average length of the videos was 25 minutes.

Video segment selection. From the data set of 64 videos, segments were systematically sampled from all videos collected during weeks two, five, and eight of the intervention study. These weeks were chosen to ensure that instruction was being observed throughout all phases of the intervention. These particular weeks were selected

because one school began recording during the first week of the study, but the other school didn't begin recording until the second week, so by beginning with the second week, the time of year in school was held constant across sites. The eighth week was chosen because it is the last common week across both sites, and the fifth week was equidistant from the 2nd and 8th weeks. Videos from these three weeks range in length from 15 minutes to 29 minutes with the average video length for these three weeks being approximately 25 minutes. Allowing for three observations of each group resulted in nine separate ten-minute segments for each interventionist, except for the interventionist who instructed two groups, for whom there were 18 ten-minute segments. The original intent of the study was to examine the use of six-minute lesson segments to measure implementation. Due to reliability issues (which will be addressed within the observation procedures section of this chapter and in the reliability section of chapter 4), the decision was made to increase the length of video segments to ten minutes.

To gather the ten-minute samples from the lesson, each video was evenly divided into three sections (beginning, middle, and end) and ten consecutive minutes were randomly-selected from each of these three sections. Random selection occurred through use of a random number generator to choose the starting point for the video segment from minute one through minute four. The following consecutive ten minutes from that starting point were used. Each segment was randomly-coded to ensure that observers were blind to segment of the lesson (beginning, middle, end) or intervention phase (2nd week, 5th week, 8th week). Due to length of some full-length videos, overlap in segments sometimes occurred. Selection of ten-minute segments was informed by the work of Pratt and Logan (2014), discussed in the previous chapter, and Giraletto and Weitzman (2002)

who were able to accurately and reliably code pre-school teacher-child language interactions shorter segments of instruction. Although Giraletto and Weitzman (2002) used slightly different segment lengths, this study and that of Pratt and Logan (2014) demonstrated that reliable and valid subsamples could be collected from an overall lesson.

Training and Observation Procedures

Training procedures. Observers were trained to use the QIDR following the same protocol as the initial study. Each observer was required to attend two 3-hour training sessions conducted by the principal investigator. During these sessions, observers were introduced to general procedures involved in observation and rating of videos, as well as the QIDR coding scale, general characteristics of the QIDR rubric, and tools to aid in note-taking during observation. Coders were then introduced to the items in the rubric in subsections. The subsections are groupings of items within the rubric that have some commonalities. The first part of rubric training involved the first four items which are related to organization and management (e.g., organization of materials, smooth transitions). The next segment of training involved the three items on the rubric most related to provision of emotional and behavioral support during instruction (i.e., positive reinforcement, response to problem behaviors, response to emotional needs). The final section of the rubric involves four items related to instructional practices (i.e., consistent wording, clear signals, modeling, and error corrections). Observers were asked to explore the differences between different levels of implementation for each of the items within a subsection and were then presented with a video segment (approximately five minutes in length) and asked to score the instruction based on only the items for the subsection being

discussed (e.g., organization, management). After independent scoring for each subsection, an opportunity for comparison with "true" scores (derived using a consensus-building approach during development of the QIDR), and discussion regarding justification for those scores, was provided. A different video segment was presented for each of the different segments of the rubric. Once all three subsections of the rubric had been presented with opportunities for practice scoring, a full-length practice video was presented and observers scored using all 19 items contained in the rubric. "True" scores were then presented to observers for comparison along with opportunities for discussion of discrepancies. Additional examples and practice were provided if observers had multiple discrepancies of more than one point off of "true" scores for practice videos. Once initial training was complete, observers coded a practice video independently. If observers were able to obtain 80% within one-point agreement with "true" scores, they were assigned their first set of video segments to score. All coders achieved this level of reliability at the onset of coding cycles.

Observation procedures. Video segments were randomly assigned to observers, with an eye toward ensuring balance across coders (i.e., one observer is not coding videos from the same interventionist, lesson segment, or intervention phase disproportionately). Thirty-six percent of the video segments were coded by two or more observers for purposes of measuring inter-rater reliability. Although other studies have relied on double-coding of a lower percentage (e.g., Pratt & Logan: 14%; Pianta, et al., 2008: <10%), these studies typically involve much larger data sets than were utilized for this study. Therefore, a higher percentage was selected to ensure reliability of observations.

Assignment of each video segment was stratified so that a different coder was randomly-assigned to each segment taken from each full-length video. Once video segment assignment had been completed, each observer received a file containing only the video segments that they would be scoring using the QIDR.

Inter-rater reliability (IRR). IRR was assessed on a randomly-selected 36% (*n*=26) of video segments. Coders were systematically assigned to distribute reliability videos evenly across coders to ensure all possible coder pairs were utilized in the analyses. Reliability checks were completed weekly during data collection to ensure that rater drift was not present and so that re-calibration could occur as needed. This process, described in more detail in Chapter IV, led to both the increase in video lengths from six to ten minutes, and the elimination of two coders when re-training and calibration efforts were unsuccessful for those two coders.

Confidentiality. Informed consent was collected from all student and instructional assistant participants at the beginning of the original *Super K* project. Additional measures were taken to protect participant confidentiality throughout this project. All observers were required to complete CITI training and sign a confidentiality agreement instructing observers to not share videos, the observation tool (i.e., QIDR), and data from the project. All observer records were collected and destroyed once analysis was complete, and videos were deleted from observer computers at the end of the project.

Experimental Design and Analytic Approach

The current study examined the relationship between QIDR scores obtained from full-length instructional sessions with scores obtained after viewing only 10-minute

segments of the same sessions. Classical test theory techniques for estimating reliability were performed on all factor scores. The factor scores within this study include group, full-length lessons, segment of the lesson (e.g., beginning, middle, end), and intervention phase (2^{nd} week, 5^{th} week, 8^{th} week). The goal when using classical test theory is to determine the degree to which variations in the conditions of measurement (e.g., different observers and different lessons) affect the consistency with which a construct is measured (Briesch et al., 2014). This theory assumes that any observed score is composed of a true score and some degree of measurement error. Since measurement of a true score is not possible, classical test theory allows the researcher to estimate the true score by determining the average score obtained across the administration of a hypothetically infinite number of parallel measurements (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Each observed score represents an attempt to estimate the true score, but the presence of some degree of random and predictable error is always assumed (Osterlind, 2006).

There are inherent limitations with using classical test theory within this study. The most notable limitation is that classical test theory only explains one general source of error, not taking into account the possible various sources of error (Brennan, 2010). For the purposes of this study, however, classical test theory is a sufficient method for estimating the reliability of measurements under the specified test conditions given that the purpose is simply to determine levels of reliability within each condition. The following section will outline the analysis methods that were used to answer each research question.

Can adequate inter-rater reliability (IRR) be obtained after observing 10 minutes of 30-minute full-length intervention lessons? Research involving the CLASS (e.g., Pianta, et al., 2008; Pratt & Logan, 2014) has used percent agreement within one point between observers. Hallgren (2012) criticized the use of this method given that percentage of agreements does not correct for agreements expected by chance and consequently overestimates the level of agreement. Because of this, the first research question was addressed using one-way random, single-measures *ICC*s with absolute agreement. Adequate inter-rater reliability was defined as an *ICC* value of .60 or above as this is a commonly-cited cutoff for a good rating of agreement based on *ICC* values (Cichetti, 1994). Recommendations by Hallgren (2012) guided the selection of this particular method. The first consideration when choosing a method for calculating IRR using *ICC*s involved how many observers would code each video segment. Although having all coders code all video segments is theoretically preferred, the time-intensive nature of carrying out this design was prohibitive (Hallgren, 2012). Therefore, all observers did not view all video segments, calling for a one-way model rather than a two-way model in which all observers would code all segments. For the purposes of this study, a subset of videos was rated by two or more randomly-selected observers (36%; *N* = 26), while the remainder were coded by randomly-selected single observers.

According to Hallgren (2012), the next consideration involves whether or not good IRR is achieved by absolute agreement or consistency in ratings (i.e., rank ordering). Within this study, the purpose was to investigate the ability of coders to provide similar ratings with each other, not whether each observer's ratings remained consistent across the study, therefore absolute agreement was necessary.

Hallgren (2012) also recommends that the unit of analysis be considered when selecting a method for measuring IRR using *ICC*s. If all videos were being coded by multiple observers with the average of their ratings being used for hypothesis testing, average-measures *ICC*s could be used. Given that this study involved a subset of videos coded by two or more observers that were meant to generalize to the videos rated by only one observer, a single-measures *ICC* was used.

Using the QIDR, what is the relationship between scores obtained watching the full lesson versus sampling ten minutes of the lesson? The second research question involved the relationship between scores obtained after observations of the full lesson versus the ten-minute samples and was calculated using Pearson product-moment bivariate correlations (Field, 2013; Miles & Banyard, 2007). Scores obtained after viewing each ten-minute segment were compared to scores obtained after viewing the corresponding full-length video to determine the strength of the relationship between the segments and the full-length videos.

The analysis for the third research question, which also examined the relationship between scores obtained in the various time segments involved a two-step process. First, general descriptives were obtained including means, standard deviations, and correlations. Next, a two-way, within-subject, repeated factors ANOVA was used to test for equivalence. Within-subject, repeated measures analysis was used because the analysis was conducted comparing repeated measures of the separate lesson segments and intervention phases for each interventionist using the same measurement tool across all conditions. Repeated measures ANOVA allows a comparison of several means obtained from the same subjects (Field, 2012). The two predictor variables were QIDR

scores obtained for the full-length lessons, the three lesson segments (beginning, middle, and end), and the three intervention phases (2nd week, 5th week, and 8th week). The dependent variable for this analysis was sum scores of the QIDR using the first nineteen items within the tool.

Which QIDR ratings (full lesson vs. 10-minute sample; beginning, middle, end; intervention phase) account for the most variance in student outcomes? This research question was answered using hierarchical linear modeling (HLM) analyses. Multi-level modeling such as HLM is appropriate due to the nested nature of the data (Luke, 2004; Field, 2012). In this study, students are nested within groups which are nested within schools. Given that each time a segment is taken from a common full lesson, indicating that independence of the data is unlikely, multi-level modeling is appropriate as it models the relationship between residuals that are dependent in nature (Field, 2012). For the purposes of this study, only two levels were considered, students and groups.

To examine reading achievement and determine the effect of group membership on student WAT scores, students are level one in the model because they are nested within groups, which are level two. The following equations were used to build the model at the student level:

Level one: $Y_{ij} = \beta_{0j} + r_{ij}$

Level two: $\beta_{0j} = \gamma_{00} + u_{0j}$

In the level one equation for this model, Y_{ij} represents the WAT score for student *i* in group *j*, β_{0j} represents the mean WAT score for group *j* , and r_{ij} represents the

residual for individual student. In level two of this model, γ_{00} is the grand mean and u_{0j} represents the variability of WAT scores between groups.

Next, to examine the effect of average QIDR scores in each condition (i.e., full-length, lesson segment or intervention phase) on student outcomes as measured by WAT, an additional multilevel model was built. A different model was used to measure the effect of each particular time segment or phase. Model 2 included the full-length QIDR scores as predictors, 3-5 addressed the lesson segments, and models 6-8 addressed intervention phases. QIDR scores for each specific time segment or phase for each group was used for separate analyses, but followed the same general equation model, entering in the scores for each time segment or phase as a separate analysis.

Level one: $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$

Level two: $\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$

In level one of this model, Y_{ij} represents the WAT score for student *i* in group *j*, β_{0j} represents the mean WAT score for group *j*, $\beta_{1j}X_{ij}$ represents the effect of the QIDR score for group *j* on WAT score for student *i* in group *j* (slope), and r_{ij} represents the error term.

In the equation in level two, β_{0j} represents the mean WAT score for group *j*, while γ_{00} is the grand mean, $\gamma_{01}W_j$ represents the effect of QIDR scores on group WAT scores (slope), and u_{0j} is the variability of reading scores between groups.

Once the relationship between QIDR scores and student outcomes had been examined separately, an analysis of the relationship between specific time segments and their ability to explain variance in student outcomes was employed. An examination of model statistics and calculations of pseudo-R^2 were used to identify the variance explained for

each model. Then results of calculations for each model were compared to determine which models explained more or less variance in student outcomes by group.

CHAPTER IV

RESULTS

Descriptive Analysis

Before performing any statistical analyses, raw student data for the Word Attack (WAT) subtest of the Woodcock-Johnson Reading Mastery Tests—Revised (WMRT-R; Woodcock, 1987), as well as group-level Quality of Intervention Delivery and Receipt (QIDR; Harn et al., 2012) observation data for each video segment and full-length video, were examined using SPSS 21.0 for Windows. Final lesson segments were ten minutes in length and were obtained from full-length intervention lessons. As discussed in Chapter 3, lesson segments were increased from six minutes to ten minutes due to issues of reliability, which will be discussed in more detail later in this chapter. Lesson segments were selected from the beginning, middle, and end of each lesson. Results regarding intervention phases will also be discussed in this chapter. Intervention phases are defined as periods of time within the entire ten-week intervention. Therefore, Phase A is instruction that occurred during the 2nd week of intervention, Phase B occurred during the 5th week of intervention, and Phase C occurred during the 8th week of intervention.

Descriptive statistics. Descriptive data for student WAT outcomes is provided in Table 2. The WAT Standard Score (SS), obtained after the intervention, was used for analysis. Only students with complete data were included in the analytic sample. Out of the 35 children in the original study, complete data was available for 31 students, so those 31 students comprised the final analytic sample for all analyses. The average WAT standard score for these at-risk kindergarten students was 99.9, with a range of 94-114.

Scores obtained from observations of the 10-minute segments and full-length segments, using the QIDR tool, were also examined. The scores consisted of the sum of the 15 items on the QIDR pertaining to instruction, as well as the four items pertaining to student response. Mean scores obtained using the QIDR tool varied, with full-length observations yielding the highest mean score (M=36.80, SD=11.33). Whereas lesson segment observations conducted at the end of the lessons had the lowest mean value and the middle lesson segment score mean was most similar to the full-length observation value. The beginning lesson segment score mean fell in between the end and middle mean value. All segment score standard deviations were quite similar ranging from 11.15 for the middle segment to 11.48 for the beginning segment. See Table 3 for complete descriptive statistics.

During each intervention phase, a QIDR score was also obtained for each group. The mean of the three segment scores within each intervention phase was also examined. In general, mean scores decreased across intervention phases, while variability increased. Mean QIDR scores ranged from 32.26 – 38.27 and standard deviations ranged from 8.00 to 13.08. Descriptive data for overall QIDR scores, lesson segments, and intervention phases is found in Table 3, while descriptive statistics presented by group and lesson segment are provided in Table 4, and descriptive statistics by group and intervention phase are provided in Table 5. For a visual representation of the lesson segment and intervention data by group, boxplots are provided in Figures 1 and 2.

Table 2

Student Outcome Descriptive Statistics

	N	Min	Max	*M*	*SD*
WJ Word Attack (SS)	31	94	114	99.9	7.47

Note. WJ= Woodcock Johnson; SS=Standard Score

Table 3

*Descriptive Statistics of Overall Quality of Intervention Delivery and Receipt Scores by Lesson Segment and Intervention Phase(*N *= 24)*

Overall	*M*	*SD*	Min	Max	*ICCs*
Beg Segments	35.75	11.48	16.00	56.00	.72
Mid Segments	36.23	11.15	15.00	52.00	.62
End Segments	35.41	11.87	9.00	50.00	.77
Full-Length	36.80	11.33	17.00	56.00	.81
Phase A	38.27	8.00	26.00	53.00	
Phase B	36.86	11.90	15.00	56.00	
Phase C	32.26	13.08	9.00	50.50	

Table 4

*Descriptive Statistics of Quality of Intervention Delivery and Receipt by Group and Lesson Segment (*N = *3)*

Group		*M*	*SD*	Min	Max
1	Beg Segment	42.67	6.11	36.00	48.00
	Mid Segment	40.43	3.60	36.80	44.00
	End Segment	48.00	2.65	45.00	50.00
	Full-Length	47.33	8.08	38.00	52.00
2	Beg Segment	44.50	3.28	41.50	48.00
	Mid Segment	39.83	4.48	37.00	45.00
	End Segment	42.50	6.76	35.50	49.00
	Full-Length	38.67	10.26	30.00	50.00
3	Beg Segment	42.33	7.77	36.00	51.00
	Mid Segment	38.83	6.45	33.50	46.00
	End Segment	45.27	5.83	41.80	52.00
	Full-Length	40.40	8.12	32.00	48.20
4	Beg Segment	37.00	3.12	34.50	40.50
	Mid Segment	39.00	6.25	34.00	46.00
	End Segment	42.33	4.51	38.00	47.00
	Full-Length	36.80	8.91	27.00	44.40
5	Beg Segment	49.33	7.02	42.00	56.00
	Mid Segment	53.17	2.75	50.50	56.00
	End Segment	47.54	2.65	44.50	49.33
	Full-Length	47.67	2.08	46.00	50.00
6	Beg Segment	26.67	0.58	26.00	27.00
	Mid Segment	20.83	5.01	16.00	26.00
	End Segment	20.11	8.00	15.00	29.33
	Full-Length	23.42	12.76	9.00	33.25
7	Beg Segment	27.00	3.61	24.00	31.00
	Mid Segment	25.00	6.00	19.00	31.00
	End Segment	25.33	4.51	21.00	30.00
	Full-Length	25.33	3.06	22.00	28.00
8	Beg Segment	20.33	4.93	17.00	26.00
	Mid Segment	26.67	9.87	20.00	38.00
	End Segment	26.33	7.37	18.00	32.00
	Full-Length	23.00	8.19	16.00	32.00

Table 5

*Descriptive Statistics of Quality of Intervention Delivery and Receipt by Group and Intervention Phase (*N = *3)*

Group		*M*	*SD*	Min	Max
1	Phase A	40.60	7.29	36.00	49.00
	Phase B	44.50	3.77	40.50	48.00
	Phase C	46.00	3.47	44.00	50.00
2	Phase A	36.17	0.76	35.50	37.00
	Phase B	48.00	2.65	45.00	50.00
	Phase C	36.83	6.53	30.00	43.00
3	Phase A	48.73	3.04	46.00	52.00
	Phase B	42.80	5.19	37.00	47.00
	Phase C	35.83	5.39	32.00	42.00
4	Phase A	38.33	4.04	34.00	42.00
	Phase B	39.00	6.25	34.00	46.00
	Phase C	37.00	9.54	27.00	46.00
5	Phase A	49.17	4.31	44.50	53.00
	Phase B	50.44	5.09	46.00	56.00
	Phase C	48.77	1.75	47.00	50.50
6	Phase A	29.53	3.63	26.00	33.25
	Phase B	21.17	6.53	15.00	28.00
	Phase C	13.67	4.04	9.00	16.00
7	Phase A	29.67	1.53	28.00	31.00
	Phase B	24.00	1.73	22.00	25.00
	Phase C	22.00	3.61	19.00	26.00
8	Phase A	34.00	3.46	32.00	38.00
	Phase B	24.00	4.36	21.00	29.00
	Phase C	18.00	2.00	16.00	20.00

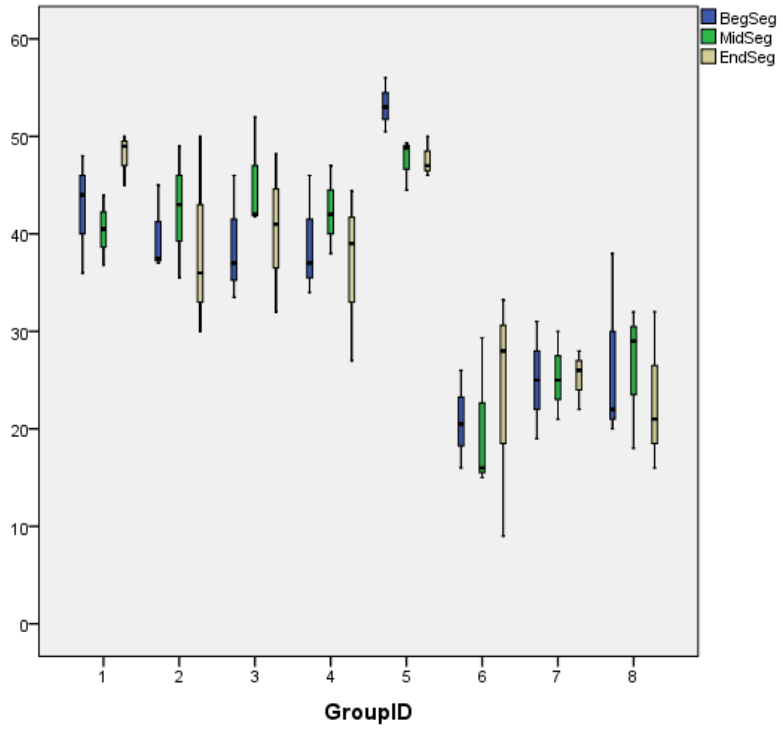

Figure 1. Boxplots of group QIDR scores by lesson segment.

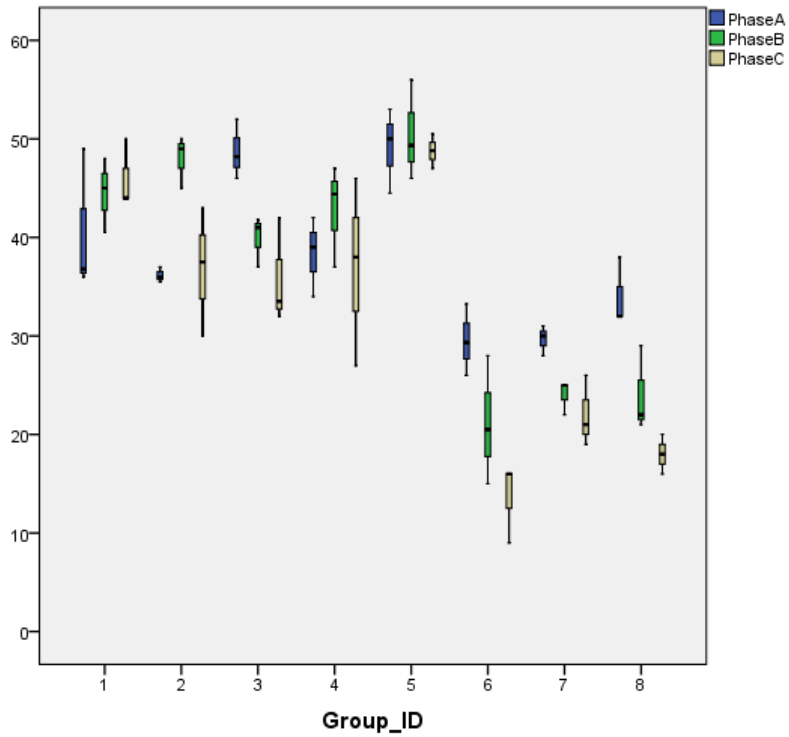

Figure 2. Boxplots of group QIDR scores by intervention phase.

Normality was examined for all variables in the data set. Although the WAT scores demonstrated distribution statistics that were within the acceptable range based on various rules of thumb (i.e., skewness +/- 1 and kurtosis +/- 2; George & Mallery, 2010; Tabachnick & Fidell, 2013) with skewness of 0.69 (*SE* = 0.42) and kurtosis of -1.18 (*SE* = 0.82), an examination of the WAT score histogram suggested that these data were positively skewed. As such, it was determined that normality could not be assumed within this data set and additional steps were taken to ensure the data were appropriate for the proposed analyses. Specifically, because fifty-eight percent of students (n=13) obtained the minimum standard score of 94 (raw score = 0), floor effects for the WAT measure were examined. Previous research has noted that it is common to see floor effects with measures of early literacy, particularly with those children who have had little or no exposure to early literacy instruction (Catts, Petscher, Schatschneider, Bridge, & Mendoza, 2009). The children included in this data set, however, had received approximately 10 weeks of literacy intervention on top of their general education literacy instruction prior to the post-test WAT measure, therefore the floor effect in this data set is likely due to the at-risk nature of the kindergarten students included in the intervention (Forbes-Spear, 2014). Forbes-Spear (2014) determined that when removing the standard scores of 94 from the data set, bivariate correlations were systematically higher, and that including the scores of 94 would actually provide a more conservative estimate of the relationships between WAT and the implementation variables that were examined.

To verify this relationship within the current study, bivariate correlations were run between the full sample of scores (n = 31) and the restricted sample with scores for students obtaining a standard score of 94 removed. These correlations are displayed in

Table 6. Similar to Forbes-Spear's (2014) findings, correlations between the full sample and the QIDR lesson segment measures are lower than those of the restricted sample, with the exception of the correlation between the end segments and the full and restricted sample, 0.50 and 0.49, respectively. It should be noted that all correlations using the full sample of WAT scores were significant at $p < .05$, with the exception of the middle segment correlation, which was not statistically significant. No correlations were statistically significant for the restricted sample, however the lack of significance within the restricted sample may be a result of the reduced sample size. As Forbes-Spear (2014) determined, the fact that the correlations using the full sample were smaller indicate that using the full sample will provide a more conservative estimate of the relationship between QIDR scores and WAT scores. Multi-level models are also more effective at accounting for violations of abnormality (Maas & Hox, 2004). For these reasons, the full sample of WAT scores were included for all analyses.

Table 6

Bivariate Correlational Analysis of Group Differences between Full and Restricted Sample

QIDR Score	Full Sample WAT_SS (*n*=31)	Restricted Sample WAT_SS (*n=13)*
Beginning Segments	0.41*	0.46
Middle Segments	0.29	0.52
End Segments	0.50**	0.49
Full Length	0.45*	0.49

Note. WAT_SS = Word Attack Standard Score; QIDR=Examining Quality of Intervention Delivery and Receipt. *$p < .05$; **$p < .01$

Data for each of the segments, as well as the full-length measures, of QIDR were also examined for skewness, kurtosis, and severe outliers. All fell within the normal distribution range, with no severe outliers, skew, or kurtosis. Bivariate scatterplots of WAT scores and QIDR scores, including lesson segment and the full-length observations, were examined and revealed no significant outliers, and no notable differences between each lesson segment and full-length measures. Bivariate scatterplots were also generated to compare segment scores with full-length measures of QIDR. These scatterplots indicated that all segments and the full-length measure had clear linear relationships.

Testing of model assumptions. Assumptions were assessed for each multi-level model by examining the final model residuals using HLM version 7.01 for Windows (Raudenbush, Bryk, & Congdon, 2013) and SPSS 21.0 for Windows. Even considering the floor effects on the outcome variable (WAT), residuals were normally distributed and independent. Residuals obtained from analyses of each lesson segment length of the QIDR were also normally distributed and independent.

Results

Research Question 1: Can adequate inter-rater reliability (IRR) be obtained after observing only 10 minutes of full-length intervention lessons? Thirty-six percent of the video segments (n = 26) were selected through stratified random sampling, controlling for both lesson segment and intervention phase, to assess inter-rater reliability. Seven coders were initially trained to observe and code video segments. As discussed in the methods section of this document (Chapter 3), the original intent of the study was to explore the use of six-minute lesson segments for scoring using the QIDR. After training, these seven coders were assigned the first set of six-minute lesson

segments for coding, and reliability was assessed. Cichetti (1994) provided the following guidelines for acceptable *ICC* ratings: values between .60 and .74 classified as good, and between .75 and 1.0 as excellent agreement. The reliability for the first week was extremely low, with an *ICC* of .06. Retraining and recalibration was attempted, but even after these efforts, an *ICC* of .20 was achieved, which was also much lower than the "good" rating suggested by Cichetti (1994). An examination of scores on specific items, as well as specific raters, was conducted to determine the source of issues of reliability. With the data collected to this point, no clear patterns emerged, suggesting that there was not a correctable problem. It was hypothesized that the multi-faceted nature of the QIDR tool may be impacting observation with six-minute segments. Although the original six-minute length was informed by the *Snippets* research (Pratt & Logan, 2014), the *Snippets* tool had a much narrower focus (i.e., looking for use of discrete reading comprehension strategies) than the QIDR, so it was hypothesized that increasing the length to account for the more complex and multi-faceted nature of the QIDR observation tool might increase the possibility of gaining a more acceptable level of agreement.

Even with this increase to ten-minute segment lengths, reliability was variable across coding weeks. After the first week of coding of ten-minute videos, the reliability achieved was just under the acceptable level of .60 (*ICC* = .53), which warranted a re-train conversation with all coders. The next week's coding elicited a much higher reliability rating (*ICC* = .80), so the third week's coding assignments were distributed. The third week also elicited a good level of agreement with an *ICC* of .64. The final week's coding assignments were then assigned and the level of agreement for this week was far below an acceptable level (*ICC* = .23), resulting in an overall reliability of .53.

This prompted an in-depth exploration of the scoring patterns of individual coders. To do this, five segments were coded by all coders, one was coded by six of the seven coders, and six were coded by different combinations of five coders. This allowed for a comparison of all possible pairs of coders. Through this comparison, it was discovered that two of the seven coders were systematically unreliable with other coders. When these two coders were eliminated, *ICCs* increased to .70 overall. For this reason, all lesson segments coded by these two coders were eliminated from the sample and randomly re-assigned to remaining coders.

The elimination of these two coders resulted in acceptable levels of reliability across the study. Final IRR was assessed using a one-way, random-effects, absolute-agreement intra-class correlation (*ICC*; McGraw & Wong, 1996) to determine the degree to which coders agreed upon ratings of lesson segments. As seen in Table 7, the resulting average *ICC* for all video segments was in the good range, *ICC* = .71, indicating that coders had moderate to high agreement. *ICC*s were also calculated by lesson segment to determine if beginning, middle, or end segments elicited higher rates of inter-rater agreement. The highest level of agreement between raters occurred within end segments (*ICC* = .77), while the lowest agreement occurred within middle segments (*ICC* = .62); however, as seen in Table 7, all segments and overall agreement fell within the good or excellent range (Cichetti, 1994). These ratings indicate that measurement error introduced by the final five independent coders was minimal, regardless of observation length and segment of the lesson, and that QIDR ratings were suitable for use in additional analyses in the present study.

Table 7

One-way, Random-effects, Absolute Agreement Intra-class Correlations for Assessing Inter-rater Agreement by Segment and Overall

Segment/Overall	Beginning (*n* = 9)	Middle (*n* = 10)	End (*n* = 7)	Overall (*n* = 26)
*ICC*s	.72	.62	.77	.71

Research Question 2: Using the QIDR, what is the relationship between scores obtained watching the full lesson versus sampling ten minutes of the lesson?

Pearson product-moment bivariate correlations (Field, 2013; Miles & Banyard, 2007) were used to calculate the relationship between observation scores obtained from full-length lessons and those obtained from ten-minute segments. In addition, lesson segment scores obtained across phases were averaged and correlated with the scores obtained from full-length lessons. Given that all observations were obtained from the same set of videos, strong correlations between scores were expected. Relationships between all segments and the full-length observations were strong, positive, and statistically significant at the $p < .01$ level. Full-length observations were most strongly correlated with beginning segments, followed by middle segments, and end segments with correlations ranging from .72 to .81. Table 8 provides an overview of correlational analyses between lesson segments and full-length observations. Correlations between intervention phase scores from lesson segments were also strongly and significantly correlated with full-length observation scores. The weakest correlation was between Phase A lesson segment scores and full length observations ($r = .77$, $p < .05$), and phases B & C lesson segment scores were similarly highly correlated with the full-length lesson

scores, (r = .94, p < .01 and r = .95, p < .01, respectively). Table 9 provides an overview of correlational analyses between intervention phases and full-length lessons.

Table 8

*Bivariate Correlations for QIDR Ratings Between Full-length Observations and Lesson Segments (*N = *24)*

Lesson Segment	1	2	3	4	5
1. Full-length Observation	-				
2. Beginning Segment	.81**	-			
3. Middle Segment	.74**	.88**	-		
4. End Segment	.72**	.82**	.84**	-	

Note. **p < .01

Table 9

*Bivariate Correlations for QIDR Ratings Between Full-length Observations and Intervention Phases (*N = *24)*

Phase	1	2	3	4	5
1. Full-length Observation	-				
2. Phase A Segment Average	0.77*	-			
3. Phase B Segment Average	0.94**	0.75*	-		
4. Phase C Segment Average	0.95**	0.80*	0.95**	-	

Note. **p < .01; *p < .05.

Research Question 3: To what extent does the relationship between QIDR ratings obtained watching the full lesson, versus sampling ten minutes of the lesson, depend on lesson segment or on intervention phase? Data were analyzed as a two-way, within-subject, repeated measures ANOVA to test for equivalence. The two within-

subjects predictor variables were the lesson segments (beginning, middle, end, and full lesson), and the intervention phase (2nd week, 5th week, 8th week, and average overall full lesson). The dependent variable was the total score (i.e., sum of the first 19 items) of the QIDR. The average of each group's full-length QIDR score was used to calculate the differences. Unadjusted *p*-values were used to evaluate within-subjects effects because the assumption of sphericity was evaluated with the Mauchly Sphericity Tests and found to be tenable for both lesson segment and intervention phase, $\chi^2(5) = 5.70, p > .05$ and $\chi^2(5) = 5.75, p > .05$, respectively. The analysis of variance results are reported in Table 10. There was not a statistically significant effect of lesson segment, nor intervention phase $F(3, 67) = 0.34$, $p = .80$, and $F(3, 21) = 2.85$, $p = .06$, respectively. Although the effect of phase was nearing statistical significance, neither lesson segment, nor intervention phase statistically significantly explained the variance in scores on the QIDR.

Table 10

One-Way, Within-subjects, Repeated Measures Analysis of Variance Summary Table for the Effects of Lesson Segment and Intervention Phase on QIDR Scores

Source	*df*	*SS*	*MS*	*F*	*p*	η^2
Within Subjects						
Lesson Segment	3	26.73	8.91	.34	.80	.02
Error within	69	1814.14	26.29			
Within subjects						
Intervention Phase	3	164.19	54.73	2.85	.06	.29
Error within	21	404.03	19.24			

Research Question 4: Which QIDR ratings (full lesson vs. 10-minute lesson segment; beginning, middle, end; intervention phase) account for the most variance in student outcomes? For these analyses, hierarchical linear modeling (HLM) was employed to analyze variance in WAT scores that could be explained through analysis of the full-length model, lesson segments (beginning, middle, and end), and intervention phases within the intervention period (2nd week, 5th week, and 8th week). Table 11 provides an overview of the model estimates for each model.

Null model. To begin, the null model was used to estimate the variance at each level, with no predictors entered into the model. This analysis indicated that there was significant variance at the student level $t(7) = 48.05$, $p < .001$, and the group level, $\chi^2(7) = 27.76, p < .001$. When *ICC*s were calculated, it was determined that 56% of the variance in WAT scores occurred at the student level, while 44% of the variance occurred between groups. These results corroborated the results found by Forbes-Spear (2014) with the same data set, and led to the same conclusion that multi-level models were the appropriate analyses, given the large variance at the group level.

Full-length QIDR measure. Next, the average of each group's full-length QIDR scores, obtained from the 2nd, 5th, and 8th weeks of instruction, was entered into the model to determine how much variance could be explained from a score obtained from observing a full-length intervention lesson. The coefficient for the full-length QIDR was not significant $t(6) = 1.63$, $p = 0.15$. Because of the small sample size, resulting in an underpowered study, it is not unexpected that this relationship was not significant. For this reason, additional parameters were examined in each model to explore how well each

lesson segment or intervention phase measure predicted student outcomes. Therefore, for each model, level two *ICC*s and level two pseudo-R^2 were calculated.

In this full-length QIDR measure model the amount of variance shifted from the null model, with 65% of the variance now at the individual level, and 35% of the variance at the group level. This level two variance was significant, $\chi^2(7) = 17.16, p < .01$, and the amount of variance explained by adding the QIDR as a level two predictor indicated that 30% of the variance at level two was accounted for by the full-length QIDR score, pseudo-R^2 = 0.30.

Lesson segment QIDR measures. Next, each lesson segment was entered into the model separately to determine if a particular lesson segment explained more or less variance in student WAT performance. When the average of each group's QIDR score was entered for each lesson segment, none of the coefficients for any of the lesson segments (beginning, middle, or end) were statistically significant, $t(6)$ = 1.41, 0.78, and 2.04, respectively, with all *p*-values greater than .05. It is important to note, however, that the end segment had a *p*-value that was approaching statistical significance, p = .09.

An examination of model statistics for each lesson segment model revealed slight shifts in the variance explained at each level, for each lesson segment. When the beginning segment was entered as the individual predictor the amount of variance at level one was 62%, and 38% at level two. This represented a shift from the null model, but the shift in variance was similar to the variance explained when the full-length QIDR scores were entered as the predictor. The amount of variance explained at each level was similar to the null model when the middle segment was entered as the individual predictor, with 55% of the variance explained at level one, and 45% of the variance explained at level

two. However, when the end segment was entered as the individual predictor, there was a more pronounced shift in variance explained when compared to the null model, with 70% of the variance explained at level one, and only 30% of the variance explained at level two.

The level two variance in each of the lesson segment models (beginning, middle, and end) was statistically significant. Model 4, in which the middle segment was entered, was significant at the $p < .001$ level, $\chi^2(6) = 23.47$. When pseudo-R^2 was calculated for this model, the amount of variance explained was negligible, pseudo-R^2 = -0.05. This finding indicates that middle segments did not provide any explanation of variance in group WAT scores. The other two segments, beginning and end, provided stronger models for predicting group WAT scores. Level two variance for beginning segments was significant, $\chi^2(6) = 19.02$, $p < .01$. By adding the beginning segment as a level two predictor, 20% of the variance was accounted for by the beginning segment QIDR score, pseudo-R^2 = .20. The model with the best fit was the one in which the end segment scores were entered as the individual predictor, $\chi^2(6) = 14.96$, $p < .01$. This model explained 45% of the variance in group level WAT scores, pseudo-R^2 = 0.45, explaining more variance in WAT scores than the model in which full-length QIDR scores were entered as the individual predictor (pseudo-R^2 = 0.30).

Model deviance decreased for each of the lesson segment models, but most markedly within the end-lesson segment. This, coupled with the higher pseudo-R^2 for the end-lesson segment, indicates that the end-lesson segment may be the strongest predictor of group differences in students' WAT scores, while the middle segment of each lesson appears to be a less effective predictor of group differences in students' WAT scores.

Intervention phase measures. Following the examination of the lesson segments as individual predictors, mean QIDR scores for intervention phases were individually entered into the model to determine if average 10-minute QIDR scores within a particular intervention phase were more or less predictive of student outcomes on the WAT. Once again, when the average of each group's QIDR scores were entered for each intervention phase (2nd week, 5th week, 8th week), none of the coefficients for any of the phases were significant, $t(6) = 1.12, 1.34, 1.36$, respectively, with all p-values greater than .05.

Level two variance for each of the intervention phases (2nd week, 5th week, 8th week) was significant, $\chi^2(6) = 21.22, 19.33$, and 19.03, respectively, $p < .05$ for all models. Variance explained at each level for each phase of intervention also shifted somewhat from the null model. When the first phase of the intervention was entered into the model, 58% was explained at level one, and 42% was explained at the group level, which was the smallest shift from the null model. The second and third phases both revealed 61% of the variance explained at the individual level, with 39% of the variance at level two.

The level two variances were significant, $p < .01$, for all three models in which phase was entered in as the predictor; however, based on the pseudo-R^2 calculations, none predicted group differences in student outcomes as well as the previous models involving beginning and end segments. For the 2nd-week phase of intervention, 8% of the variance could be explained by QIDR scores for the phase, pseudo-$R^2 = 0.08$, while during the 5th week of the intervention, 17% of the variance was explained by QIDR scores, pseudo-$R^2 = 0.17$, and during the 8th week of intervention, 19% of the variance

was explained by QIDR scores, pseudo-R^2 = 0. 19. These findings indicate that scores obtained later in the intervention explained the most level 2 variance

Table 11

Fixed and Random Effects Estimates Models WAT Posttest Scores by Lesson Segment and Intervention Phase

Parameter	Model 1 (Null)	Model 2 (Full-length)	Model 3 (Beg Segment)	Model 4 (Mid Segment)	Model 5 (End Segment)	Model 6 (Phase A)	Model 7 (Phase B)	Model 8 (Phase C)
Fixed Effects								
Intercept	100.37*** (2.09)	89.65*** (6.79)	100.50*** (1.93)	100.45*** (2.13)	100.56*** (1.57)	100.47*** (2.02)	100.50*** (1.95)	100.49*** (1.94)
QIDR Score		0.30 (0.18)	0.19 (0.20)	0.17 (0.21)	0.32 (0.14)	0.33 (0.28)	0.23 (0.17)	0.22 (0.16)
Random Effects								
Group (intercept)	25.98***	18.17**	20.81**	27.39***	14.21*	23.74**	21.45**	21.15**
Student residual	33.15	33.59	33.45	33.31	33.65	33.36	33.46	33.49
Model Statistics								
ICC—Level 1	.5606	.6489	.6167	.5488	.7031	.5842	.6093	.6129
ICC—Level 2	.4394	.3511	.3833	.4512	.2969	.4158	.3907	.3871
Pseudo R^2								
Level 2		0.3004	0.1988	-0.0542	0.4534	0.0862	0.1743	0.1858
Level 1		-0.0135	-0.0092	-0.0050	-0.0152	-0.0064	-0.0094	-0.0104
Deviance	203.23	200.50	200.90	202.03	199.75	200.76	201.32	201.46
Parameters	2	2	2	2	2	2	2	2
Deviance Change	--	-2.73	-2.33	-1.20	-3.48	-2.47	-1.91	-1.77

Note. **Parentheses denote standard errors. Level two predictors are group centered. *$p < .05$, **$p < .01$, ***$p < .001$**

CHAPTER V

DISCUSSION

Implementation science indicates that quality of implementation can only be improved through frequent feedback to interventionists/teachers (Fixsen, Blase, Metz, & Van Dyke, 2013). While multiple observation tools have been developed demonstrating the relation of specific instructional practices and student outcomes in general education (e.g., La Paro, Pianta, & Stuhlman, 2004), few have been developed for use in monitoring special education or small group interventions (Johnson & Semmelroth, 2013). Furthermore, current tools often require extended observation periods (e.g., over 60 minutes, multiple observations across days, etc.) that limit the practicality of use in schools. Additionally, there is a limited understanding of how much of a lesson needs to be observed to determine overall quality. Practitioners need tools that are reliable, efficient, and target key intervention instructional practices that are related to improved student outcomes. This study examines the use of an observation tool, the *Quality of Intervention Delivery and Receipt* (*QIDR;* Harn, Forbes-Spear, Fritz, & Berg, 2011), an implementation measure specifically designed for monitoring small group intervention. Prior efforts have demonstrated that the QIDR correlates with other commonly used measures (i.e., CLASS, and opportunities to respond) and accounts for significant variance in student outcomes (Forbes-Spear, 2014). The focus of this study was to examine issues related to how long an observation needs to be, as well as how scores from observations conducted during specific portions of a lesson or intervention are related to student outcomes using the same data set.

This study addressed the following research questions:

1) Can adequate inter-rater reliability (IRR) be obtained after observing 10 minutes of full-length intervention lessons?
2) Using the QIDR, what is the relationship between scores obtained watching the full lesson versus sampling ten minutes of the lesson?
3) To what extent does the relationship between QIDR ratings obtained watching the full lesson, versus sampling ten minutes of the lesson, depend on time segment of the lesson (i.e., beginning, middle, end) or on phase within the intervention (i.e., 2nd week, 5th week, 8th week)? In other words, are correlations between the ratings systematically stronger or weaker based on lesson segment or intervention phase?
4) Which QIDR ratings (full lesson vs. ten-minute sample; beginning, middle, end; intervention phase) account for the most variance in student outcomes?

The initial section of this chapter will relate findings of this study to prior research, and then the chapter will conclude with a discussion on implications for future research and practice.

Primary Findings

Inter-rater reliability. To answer the first research question, observers were randomly assigned to code lesson segments using the QIDR. Acceptable inter-rater reliability (IRR) was achieved when using ten-minute observations (*ICC* = .71). While the intra-class correlations (*ICC*s) obtained from the lesson segments was not as high as those obtained from observations of the full-length videos (*ICC* = .81) in the original study, the reliability ratings for each of the segments fell within the good range for

agreement (*ICC* > .60), with the reliability for end segments being in the excellent range (*ICC* = .77; Cichetti, 1994). These results indicate that individuals can be trained to reliably measure implementation on a multifaceted tool (i.e., QIDR) using a 10-minute segment, and that this measure of implementation is highly correlated with the score obtained from watching the entire lesson (Forbes-Spear, 2014). These findings are similar to *Snippets* research (Pratt & Logan, 2014), which found that the *Snippets* tool could reliably measure the use of specific comprehension instructional strategies within a reading lesson. A MET follow-up study also found that the reliability of 15-minute observations (representing 33% of full-length lessons) was comparable to full length lessons in general education classrooms when using the Framework for Teaching (FfT; Danielson, 1996), also a multifaceted observation tool (Ho & Kane, 2013). For instructional coaches and administrators in schools, these findings suggest that brief observations are highly related to what they would see if they had the opportunity to watch an overall lesson. Knowing this may actually encourage coaches/administrators to conduct more frequent observations to identify interventionists that may need additional professional development.

It should be noted, however, that the process of reaching an acceptable level of reliability using ten-minute segments was more challenging than it was to achieve adequate reliability for full-length QIDR observations in the previous study (Forbes Spear, 2014). The challenges in gaining reliability may have arisen from the length of the lesson segments, the multifaceted nature of the QIDR measure, and/or individual coder characteristics. More research is necessary to determine the specific sources of variance in reliability, and each is discussed below.

Lesson segment length. The original intent of this study (see chapters III and IV) was to use six-minute segments similar to Pratt and Logan (2014), but achieving reliability with segments of that length proved difficult. One reason reliability may not have been as challenging within the Pratt and Logan study was that the *Snippets* tool was measuring the presence or absence of much more discrete comprehension strategies rather than multiple elements of implementation on the QIDR. Therefore, the decision was made to increase the length of the lesson segments to ten minutes to determine if additional time would allow for more opportunities for coders to observe various implementation and response behaviors.

Multifaceted nature of QIDR. As found in other studies, the cognitive load required to attend to multiple dimensions during an observation may have affected observer reliability (Jerald, 2012; Joe, McClellan, & Holtzman, 2016). During the large-scale Measures of Effective Teaching (MET; Bill & Melinda Gates Foundation) project, a follow-up study was conducted to determine if complexity of an instrument had an effect on inter-rater reliability scores of segments of full-length lessons in a general education classroom (Joe, McClellan, & Holtzman, 2016). To do this, researchers compared reliability achieved when observers used only a subset of items of the FfT (Danielson, 1996), with reliability achieved when using all elements of the FfT. Lengths of observations ranged from 22 to 30 minutes and reflected approximately half of a full lesson. Researchers found that inter-rater reliability decreased significantly when observers were required to use all items within each of the observation instruments compared to when they used only a subset of items (Joe, McClellan, & Holtzman, 2016). The findings of this MET study support the notion that reliability issues within the

current study may have been a function of the complex, multifaceted nature of the tool. Whether the multifaceted nature of the QIDR is more influential on inter-rater reliability than the length of the lesson segments cannot be determined from this study, but is worth considering. For example, with the QIDR, observers were trained to watch for numerous elements of instruction within a shortened lesson that may not have occurred during that sample, such as emotional responsiveness, partner opportunity to respond, or teacher responding appropriately to problem behavior. While these instructional behaviors are important, their frequency of use is dependent on the context of the situation and may have impacted reliability in the current study.

Coder characteristics. Individual coder issues may have also affected reliability within this study. Two coders who had difficulty with reliability were removed from the original pool of coders (see Chapters III and IV for additional information regarding this removal). The reasons for their difficulties with attaining reliability are unclear, but some possibilities are discussed here.

Researchers have determined that multiple factors can affect coder accuracy (Repp, Nieminen, Linger & Brusca, 1988) including the setting of the interventions, complexity of the observation tool, and observer bias. Expectations of subject performance, or bias, may affect the observer's ability to accurately score an observation (Repp et al., 1988). In the case of this study, observers were aware of the general background of the interventionists and one of the eliminated coders may have been susceptible to closely identifying and sympathizing with the situations being observed, making that coder more likely to score the interventionists more leniently. This coder had multiple years of experience in both general education and intervention settings. During

training and calibration activities, this coder often commented that she understood the actions of the interventionist ("and may have reacted in the same way") even when the interventionist being scored was exhibiting less-than-desirable implementation behaviors. This coder seemed to rely more on emotional reactions and her own beliefs about teaching than actual element descriptors within the QIDR.

The second coder who had difficulty with reliability was an English language learner with multiple years of teaching experience outside of the United States (U.S.), but no experience teaching or observing instruction in the U.S. Experience differences may have presented some bias within this study. It is possible that different expectations regarding teacher and student behavior may have prompted this coder to score interventionists more leniently. In addition, nuances in the rubric language may have made interpretation of the rubric more difficult for this coder.

Interestingly, a post-hoc analysis was conducted on the reliability of the initial set of six-minute lesson segments with the same two coders eliminated. Again, it was found that reliability increased within that sample of observations, and reached an acceptable level of agreement on the two six-minute lesson segments assigned after re-training (*ICC* = .69). This may indicate that the largest issue impacting reliability was individual coder characteristics, rather than segment length or the multifaceted nature of the tool. Future studies should continue to investigate short segment lengths to increase efficiency further.

Relationship between lesson segment and full-length QIDR scores. To examine the second research question, QIDR scores from the full-length lessons were compared with QIDR scores obtained from lesson segments of the same lessons and average phase scores across the 10-week intervention. Results indicated that all segments

were strongly correlated to the full lesson ($r > .70$), with beginning lesson segments most strongly correlated with the full-length scores ($r = .81$). This finding was similar to a MET study comparing FfT scores from the first 15 minutes of a lesson with scores obtained for the whole lesson (Ho & Kane, 2013). Ho and Kane found there was little difference between the 15-minute score and the full-length observation. Correlations obtained comparing lesson segment phase scores with full-length phase scores were also strong and significant, ($r > .77$), with the lesson segment scores from phase C being most strongly correlated with full-length lesson phase scores ($r = .95$, $p < .01$).

These findings indicate that an observer can get a similar measure of implementation using the QIDR regardless of whether you watch a whole lesson, or any10-minute segment. The correlations between overall phase scores also indicate that a similar measure can be obtained regardless of phase within the intervention. The reason this result was attainable may relate to the targeted nature of the QIDR and the fact that it was developed to identify the use of very specific elements of explicit instruction found in the intervention programs used in this study. Fixsen (2013) posits that assessment systems that directly relate to the philosophy and critical elements of a program or practice can provide opportunities for repeated assessment and feedback. Hill and Grossman (2013) contend that observation tools that can provide specific feedback that can be readily implemented are more successful in improving instructional quality. The findings of the current study indicate that the nature of the QIDR may allow for flexibility in terms of what segment of a lesson can be observed, while also providing enough specificity in key instructional elements to guide discussion and feedback for interventionists that can improve instruction and student outcomes in an efficient manner.

Relationship between scores obtained during various lesson segments and intervention phases. To further explore the relationship between scores obtained during the various lesson segments and intervention phases, a two-way, within-subject, repeated measures ANOVA was used to determine the equivalency of the scores. Results indicated that there was no statistically significant difference between scores obtained in full-length observations and those generated during specific lesson segments or phases of intervention. This finding further supports the notion that any time segment will provide you with a similar measure of implementation to provide support for coaches/administrators to conduct implementation checks as needed (e.g., when student response is limited, prior observation showed low scores, etc.) rather than in a procedural manner (e.g., once a year/quarter).

Related to being responsive to monitoring implementation, an interesting, yet non-statistically significant finding was noted related to the phase of the intervention. As observed in Figure 2 (pg. 65) mean QIDR scores for all interventionists decreased across phases within this study. Others have found that observers tend to rate more harshly across time (Casabianca, Lockwood, & McCaffrey, 2015; Congdon & McQueen, 2000); however, the design of the current study should have corrected for this phenomenon because coders were blind to lesson time and segments were randomly assigned to coders, controlling for segment and phase.

When scores were closely examined, it was found that three of the four interventionists whose scores decreased most drastically across phases also had the lowest mean QIDR scores in Phase A. In general, if an interventionist's average score was above 35 during Phase A, the scores for subsequent phases also remained at or above

35. Those whose initial scores were below 35 in Phase A, showed declining scores across remaining phases. This finding may indicate that those interventionists who are most in need of support at the beginning of the intervention period will have continued decreases in instructional quality without feedback and coaching supports. This issue will be further discussed within the implications section of this chapter.

It is important to note, however, that the changes in scores reflect only three points in time within the intervention (2nd, 5th, and 8th week) and may not be representative of scores occurring across entire intervention phases. Forbes-Spear (2014) included all weekly full-length lesson measures of the QIDR throughout the 10-week intervention period within her study and found that there were no significant changes across time, on average. It was also noted within the Forbes-Spear (2014) study that QIDR scores were variable across the entire intervention period. Therefore, the findings of this study should be approached with caution, and it may be necessary to get multiple measures across time to get a more accurate measure of overall implementation within a specific phase of an intervention.

Association between QIDR and student outcomes. To address the final research question, multi-level modeling was used to determine if scores obtained using the QIDR were predictive of student outcomes. Due to the small sample size and issues with a floor effect on the DV (see discussion in the limitation section), these results also need to be interpreted with caution, and are considered exploratory in nature.

Scores on the QIDR, regardless of lesson segment or phase, were not significant predictors of group differences in student outcomes. However, when model statistics were examined for the full-length scores, each lesson segment, and each intervention

phase, the findings indicated that there were differences in variance explained with each predictor. The model using full-length lesson QIDR scores as predictors explained a substantial amount of variance in WAT scores at the group level (30%; pseudo-R^2 = 0.30). This was slightly different from the relationship found in Forbes-Spear's (2014) study that found scores from the QIDR accounted for 36% (pseudo-R^2 = 0.36) of the variance in WAT scores. The differences in variance explained may be attributed to differences in the two studies. In the current study, the 15 items that address instructional elements, as well as the four items related to student response were included for analysis, while Forbes-Spear omitted the student response items from her analysis. In addition, Forbes-Spear used QIDR scores across all weeks of intervention, where this study employed QIDR scores obtained from the 2nd, 5th, and 8th week of the intervention period.

Of all predictors, QIDR scores for end lesson segments accounted for the most variance in WAT scores at the group level (45%; pseudo-R^2 = 0.45). This indicates that QIDR scores obtained while observing the end of a lesson may be the best predictor of student outcomes. The fact that end segments actually explained more variance than did a full-length lesson suggests that there is some element of instruction or student response that is or is not occurring at the end of a lesson that may be key to impacting student outcomes. Notably, the end lesson segments also averaged the lowest QIDR scores across all interventionists and had the most variability, ranging from 9 to 50. The lower scores and large variability of implementation during the end segments of lessons may help to explain the differences in outcomes within groups. It is possible that those interventionists most skilled at teaching, including sustaining student engagement,

throughout the entire lesson, are more likely to have better student outcomes than those who do not possess the same skills.

While end lesson segments explained the most variance in student outcomes, beginning segments explained 20% of the variance in group WAT outcomes (pseudo-R^2 = 0.20) and middle segments provided no explanation of variance in WAT scores at the group level, making middle segments a very poor predictor of student outcomes (pseudo-R^2 = -0.05). Although previously discussed findings revealed that similar scores of implementation were obtained across segments, given the differences in variance explained across lesson segments, the best choice for an administrator or coach may be to observe the end of the lesson. The feedback provided from these end-of-lesson observations could better assist the interventionist in improving instruction in such a way as to sustain elements of quality implementation throughout the entire lesson, thus impacting student outcomes most profoundly.

The phase of intervention also revealed an interesting pattern regarding the variance explained across the models. While the variance explained across the phases was not as substantial as most of the models with lesson segments or the full-length scores as predictors, it appeared that the QIDR scores as a predictor became stronger across the intervention phases, with the final intervention phase explaining the most variance in WAT scores (19%) at the group level. This finding indicates that the QIDR score received by interventionists closest to the time of post-test may be more predictive of student outcomes than other phases of the intervention. Interestingly, when looking across scores, the average QIDR score obtained from intervention phases decreased across time, while variability in scores increased. The schools involved in this study had

multiple years of experience providing tiered supports for all students in a fully implemented RTI model. The interventionists included in this study had been previously trained in intervention delivery and had multiple years of experience using the intervention programs involved in the study, but the schools did not provide formal ongoing support and coaching. The finding that overall instructional quality decreased across phases, coupled with the final phase of the intervention explaining the most variance in student outcomes, provides additional support for ensuring that interventionists are provided frequent feedback and supports that will improve, rather than deteriorate, implementation across the intervention period (Fixsen, Blase, Metz, & Van Dyke, 2013; Hill & Grossman, 2013; Pianta, Mashburn, Downer, Hamre & Justice, 2008).

Limitations

Several limitations within this study must be considered with these findings, and may help to inform future research.

Sample size. First, given that only 31 students were nested within eight groups, the small sample size included in this study contributed to the underpowered nature of the study, specifically in consideration of the relationship between implementation and student outcomes. The insufficient power within this study makes it difficult to identify statistically significant effects (Maxwell, 2004) and may increase Type II error.

Student outcome measure. Another limitation within this study involved the use of the Woodcock-Johnson Word Attack Subtest (WAT; WMRT-R; Woodcock, 1987) as the outcome measure. Because of the developmental age and at-risk nature of the students within this study, the WAT was not sensitive enough to detect individual differences in

student reading outcomes. The WAT does target the skills being taught in the interventions, however, a more sensitive curriculum-based measure such as DIBELS or easyCBM may have more accurately captured the differences across students.

Lesson segment numbers. An additional limitation should be noted regarding the number of lesson segments included in the study. While 72 lesson segments may have been adequate for answering the research questions related to reliability and the relationship to full-length lesson scores, as the analyses began to explore the relation of lesson segment and intervention phase at the group level, the *n* needed to accurately answer some of the questions may have been too small. For instance, the 72 lesson segments were derived from only 24 full-length lessons, which represented only three lessons from each group. Therefore, when lesson segments were examined, only nine lesson segments were analyzed for each instructional group, accounting for only three beginning, middle, and end lesson segments per group. The study also only took into account three weeks out of the ten-week intervention period. Therefore, the average of one beginning, one middle, and one end lesson segment (all from the same full-length lesson) comprised the score for the intervention phase. Considering the multifaceted nature of the QIDR tool, as well as the variability across time found by Forbes-Spear (2014) in an earlier study, a more accurate measure of implementation within the phase may have been achieved with multiple lesson measures rather than being derived from only one lesson within the intervention phase. For this reason, some findings, particularly those involving phase and specific lesson segments by group, should be approached with some caution.

Observer reliability. The difficulty with achieving reliability among observers provided insight into aspects of observation that may impact reliability, but also provided some limitations that must be noted. It is important to point out that a select pool of coders was necessary to achieve reliability and that two coders had to be eliminated. Further discussion of this issue will be addressed in the implications section of this chapter.

Implications

The overall purpose of this study has a practical focus. Providing teachers with regular observations and feedback has been found to improve student outcomes when incorporated into a responsive instructional cycle (Fixsen, 2013; Pianta, et al., 2008). The focus of the current study was to determine if the QIDR could be used to measure implementation efficiently so that it might be effectively used for providing frequent feedback to improve intervention instruction.

Reliability can be demonstrated with abbreviated observation. The issue of achieving reliability with these abbreviated, ten-minute observations has important implications for both research and practice. Observation and feedback is only useful when it is also accurate and provides specific feedback that can improve instruction (Hill & Grossman, 2013). If administrators and coaches have the ability to visit classrooms to perform short observations using a tool that can inform feedback, they may be more likely to perform these observations on a more regular basis. This frequent observation and feedback loop is especially essential for interventionists who demonstrate weaker skills in early observations. Results of this study indicated that those interventionists who had lower QIDR scores at the beginning of the study, continued to have lower scores

throughout. Therefore, shorter observations could allow the administrator or coach to visit those interventionists most in need of feedback on a much more regular basis, increasing the likelihood that instruction will improve over time.

Challenges in achieving reliability in school settings. Providing specific, targeted feedback using a multifaceted tool such as the QIDR may present unique challenges for training and ongoing support for observers (Jones, Reid, & Patterson, 1975; Taplin & Reid, 1973). Therefore, providing adequate training, as well as frequent calibration checks, is vital for establishing and maintaining reliability within school settings.

The goal of initial training should be to ensure that observers adopt a view of teaching that is consistent with the tool being used for measurement (Bell, et al., 2016). Initial training using the QIDR must include a thorough explanation of instructional components most important for improving student outcomes within intervention settings so coaches and administrators are able to complete observations that are devoid of their own personal biases on instruction. To maintain measures of implementation that are reliable and aligned with the intent of the tool, it is also necessary to provide regular check-ins across time to ensure that coaches and administrators are continuing to provide accurate measures of implementation and student response.

The fact that two of the seven coders within this study were found to have issues with reliability indicates that having only one or two observers within a school may be problematic if the biases of the coder or coders prevent them from providing accurate assessment and feedback to interventionists. Calibration against a set of "master" scores (i.e., scores obtained and agreed upon by a group of experienced coders) initially and at

multiple times throughout a study may be necessary in order to ensure that particular coders are, and remain, reliable. Given the difficulty with achieving reliability with the original seven coders, observer characteristics may also be an important consideration. It may be necessary for researchers and administrators to carefully select observers who can provide objective measures of instructional quality. One MET study found that a survey of teacher beliefs could predict which teachers were most likely to be successfully trained to use the CLASS observation tool reliably (Ho & Kane, 2013). Developing a system for screening coders prior to initial training may help to conserve training resources by identifying those not likely to be reliable coders.

Equivalence of implementation regardless of lesson segment. The finding that there was no significant difference in the measures of implementation across lesson segments or intervention phases, as well as its strong correlation to the full-length lesson, provides a great deal of flexibility for observers in school settings. The knowledge that time of observation has little effect on measures of implementation, along with the earlier finding of reliability across shorter segments of lessons, allows administrators and coaches the ability to fit observations and feedback into busy schedules at their convenience. Shorter observations, with the added benefit of scheduling flexibility, may mean that more frequent observations and feedback can be provided to interventionists, thus allowing greater opportunity to improve instruction and subsequent student outcomes (Hill & Grossman, 2013; Jarald, 2012).

Although there were limitations in regards to the number of segments included in each phase, the findings regarding implementation across intervention phase may also be very important for maximizing resources for instructional supports within a school.

Similar to the idea of tiered instructional supports used in RTI, it may be possible to tailor the frequency and intensity of observation and feedback depending upon the needs of each interventionist. Myers, Simonsen, and Sugai (2011) used this approach with teachers implementing elements of a system of Positive Behavior Intervention and Supports (PBIS). Those teachers who were nonresponsive to schoolwide PBIS training (tier 1) in using specific and contingent praise were offered targeted training supports (tier 2) followed by more individualized training for those who were not responsive to targeted training (tier 3). Through their investigation, they found that all teachers benefited from additional supports, but that these supports were differential in need, meaning that some teachers required more intensive supports before a change in their behavior was observed. This approach could be applied using the QIDR, or other implementation tools, as well. Within this study, interventionists who scored high (above 35) in the first intervention phase, maintained high implementation across the intervention, so it may be possible to provide less frequent observations and less intensive coaching supports for them. In contrast, interventionists scoring below 35 during the initial intervention phase may require much more frequent coaching and feedback to ensure that implementation not only improves, but doesn't worsen across time.

Implementation is related to student outcomes. The elements of explicit instruction in which the QIDR were based have been shown to impact student outcomes and emphasize explicit, intensive, and supportive instructional methods (Gersten et al., 1997; Torgesen, 2002; Swanson, 1999). Although the small sample size, coupled with the floor effect in the outcome measure, limited the ability to definitively say that QIDR scores are predictive of student outcomes, the variance explained with each of the models

indicates that the QIDR may be effective for this purpose. In addition, the variance explained by QIDR scores within end segments indicates that instructional behaviors at the end of a lesson may be particularly impacting student outcomes. Although previously discussed findings indicated that observing an intervention lesson at any time can give a reliable measure of implementation, the considerable additional variance explained with end lesson segments might provide further guidance on optimal observation times, if the opportunity to choose an observation time is available. The ability of an interventionist to sustain high levels of implementation across an entire lesson may be the best predictor of student outcomes. In addition, the feedback that is provided based on the end lesson segment observations may be able to better target the skills most in need of improvement for that interventionist.

Future Research

The most challenging aspect of this study involved the reliability of coders. Future research needs to address training methods and how to guard against observer bias when training observers to use a multifaceted tool such as the QIDR. In addition, investigation into the observer characteristics that are optimal for use in both research contexts and school-based contexts is important. Understanding what traits are essential in observers may help future researchers to avoid reliability issues, and could optimize the utility of the QIDR as a tool for providing useful feedback to interventionists in school settings. Future research is necessary to determine if there is a screening measure that could be used to determine optimal coder characteristics with a tool such as the QIDR, similar to what was done with the CLASS (Ho & Kane, 2013). Finally, given the post-hoc analysis which revealed that a smaller subset of coders was able to achieve

reliability with six-minute segments, additional research investigating the possibility of using shorter segments is necessary.

Future research should also address the elements of instruction and student response within the QIDR tool that might overlap in the construct being measured. Reducing the number of elements that observers are discerning may help to increase reliability by reducing cognitive load required by observers (Joe, McClellan, & Holtzman, 2016). One possible way to decrease elements may be to combine certain elements. For example, potentially combining the "modulating lesson pacing" element with the "teacher ensuring students are firm on content" element may be possible. Both are addressing the interventionist's ability to adjust instruction based on student performance, so only one element to that effect may be necessary to capture this construct. In addition, there are two elements within the instruction portion of the QIDR that can be scored based on student response. The first item, "Teacher is familiar with the lesson," discusses teacher fluency with lesson formats, but also includes an element regarding whether or not students follow procedures. The other item, "Teacher expectations are clearly communicated and understood by students," states that either "the teacher explicitly reviews expectations, or it is clear expectations have been taught because all students demonstrate knowledge of expectations for behavior and academic routines, and meet or exceed expectations." These items seem to overlap with three items found on the group student response rubric, which address whether or not students demonstrate behaviors consistent with knowledge of expectations for routines, on-task behavior, and following directions. This could potentially reduce the number of items

necessary for scoring by three elements, thus giving the observer fewer elements to consider and impact reliability.

Given the practical intent of the current study, it would be remiss not to suggest the need for research to elucidate the utility of the QIDR for providing feedback to improve instruction. Research needs to be conducted that would determine if coaching using the QIDR as a prompt for guiding discussion and feedback with interventionists, did, in fact, improve implementation over time, and then, if improved implementation also resulted in improved student outcomes.

Conclusions

Quality instruction is especially important for students who are at-risk for failure. Unfortunately, due to limited resources in schools, interventionists providing instruction for these students are often the least likely to receive ongoing supports to ensure high-quality implementation of interventions (Al-Otaiba, Wagner, & Miller, 2014). Following the lead of RtI, a responsive instructional cycle can be used to provide the needed supports to interventionists in such a way that resources can be maximized for those requiring the most intensive supports. For those interventionists requiring more targeted and frequent supports, administrators and coaches must utilize tools that can help them to provide support and feedback that is useful in improving instruction and student outcomes on a more regular basis (Myers, Simonsen, & Sugai, 2011). The challenge is in creating a tool that can accurately and reliably measure implementation, provide enough information to guide specific, targeted feedback to improve instruction, but be streamlined enough that it can be used for frequent observation and feedback (Fixsen, et al., 2013; Hill & Grossman, 2013). Findings from this study provide initial support for

the use of the QIDR as a tool that meets these criteria. While additional research is needed to fine-tune the QIDR and confirm its utility as a coaching tool, the current study indicates that shorter, more frequent observations are feasible and that there is promise in the efficient use of the QIDR for just such a purpose.

APPENDIX

Quality of Intervention Delivery and Receipt

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
a) Teacher is familiar with the lesson (e.g., it is evident that teacher has previewed the lesson and demonstrates fluency with the formats and lesson activities).	Teacher does not demonstrate fluency with formats and lesson activities and students do not follow the procedures.	Teacher occasionally demonstrates fluency with formats and lesson activities and students only sometimes follow the procedures.	Teacher typically demonstrates fluency with formats and lesson activities and most students typically follow the procedures.	Teacher consistently demonstrates fluency with formats and lesson activities and all students consistently follow the procedures.
b) Instructional materials are organized (e.g., instructional materials are prepped before starting the lesson including worksheets, pencils for easy distribution; organization supports rather than detracts from effective instruction, smooth transitions, etc.).	Instructional materials are not organized.	Instructional materials are partially organized.	Instructional materials are completely organized.	All instructional materials are organized specifically by lesson or student name.
c) Transitions between activities are efficient and smooth (e.g., well-established routines are in place, "teacher talk" is minor between lesson components, less than 1-2 minutes). Excluding factors outside teacher control such as fire drill.	Teacher does not implement well-established routines to minimize interruptions. (e.g., transitions often take longer than 2 minutes, excluding outside factors).	Teacher occasionally implements well-established routines to minimize interruptions but "Teacher Talk" may occur, **or** transitions are inconsistent (e.g., transitions occasionally take longer than 2 minutes, excluding outside factors).	Teacher implements well-established routines to minimize interruptions. "Teacher talk" between transitions is minimal (e.g., transitions typically take less than 1-2 minutes, excluding outside factors).	Teacher implements well-established routines to minimize interruptions. All transitions consistently occur and activities flow nearly seamlessly (e.g., transitions consistently take about a minute excluding outside factors).
d) Teacher expectations are clearly communicated and understood by students (e.g., teacher reviews academic and behavior expectations, uses clearly established routines, precorrects for challenging activities, etc.).	Teacher does not explicitly state expectations and students do not demonstrate knowledge of expectations for behavior and academic routines.	Teacher states expectations but students only occasionally demonstrate knowledge of expectations for behavior and academic routines.	Teacher explicitly reviews expectations **or** it is clear expectations have been taught because most students typically demonstrate knowledge of expectations for behavior and academic routines.	Teacher explicitly reviews expectations **or** it is clear expectations have been taught because all students consistently demonstrate knowledge of expectations for behavior and academic routines and meet or exceed those expectations.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
e) Teacher positively reinforces correct responses and behavior as appropriate (group and individual) (e.g., teacher inserts affirmations, specific praise, and confirmations either overtly or in an unobtrusive way).	Teacher does not use positive reinforcement to reinforce correct responses and appropriate behavior through verbal and nonverbal feedback when appropriate.	Teacher occasionally uses positive reinforcement to reinforce correct responses and appropriate behavior through verbal and nonverbal feedback when appropriate.	Teacher typically uses targeted positive reinforcement (specific and general) to reinforce correct responses and appropriate behavior through verbal and nonverbal feedback when appropriate	Teacher consistently and effectively uses positive reinforcement (specific and general, individual and group) to reinforce correct responses and appropriate behavior through verbal and nonverbal feedback when appropriate.
f) Teacher appropriately responds to problem behaviors (e.g., including off task; emphasizes success while providing descriptive, corrective feedback; positively reinforces to get students back on track).	Teacher does not appropriately respond to problem behavior across multiple students. Teacher primarily provides negative feedback or ignores problem behavior for extended period of time (resulting in limited student participation, e.g., more than 20% of activity).	Teacher sometimes appropriately responds to problem behavior. Teacher provides some positive or corrective feedback but does not regularly emphasize success. Teacher may have difficulty consistently responding to one student's problem behavior but sometimes responds appropriately to other students.	Teacher typically responds appropriately to problem behavior by emphasizing success and providing neutral corrective feedback for most students. ***Or*** no problem behavior occurs during the instruction.	Teacher consistently responds appropriately to problem behavior by emphasizing success and providing descriptive corrective feedback as needed for all students. For example, teacher "catches" students engaging in appropriate behavior and provides descriptive positive feedback to encourage appropriate behavior.
g) Teacher is responsive to the emotional needs of the students (e.g., teacher connects not only academically but personally to students, calls them by name, jokes with them, asks about their day, etc.).	Teacher provides limited/no positive feedback, may use sarcasm, and is unresponsive/unaware of students' emotional needs.	Teacher is generally neutral, may provide positive feedback but is directed toward academic content (i.e., no demonstration of being aware of students' emotional needs).	Teacher is typically positive, responsive and aware of most students' emotional needs. Teacher greets students by name, makes students feel welcome, respects their individuality, makes an effort to make a connection, and appears to enjoy students.	Teacher is consistently very positive, responsive and aware of all students' emotional needs. Teacher greets students by name, makes students feel welcome, respects their individuality, makes an effort to make a connection, and appears to enjoy students.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
h) Teacher uses clear and consistent lesson wording (e.g., using the exact wording or a close approximation of the language of the program consistently across activities).	Teacher does not use guide including script or format. Wording is inconsistent, and there appears to be excessive "teacher talk".	Teacher partially uses guide including script or format. Wording is sometimes consistent (during particular activities or instructional components).	Teacher typically uses guide including script or format. Wording is consistent and directions are clear and easy to follow across activities.	Teacher consistently uses guide including script or format. Wording is always consistent, and directions are clear and easy to follow across all activities.
i) Teacher uses clear and consistent auditory or visual signals (e.g., it is clear to students when and how to respond appropriately during individual, partner and group responses, across all components of lesson).	Teacher does not use clear auditory or visual signals to ensure students respond appropriately.	Teacher occasionally uses clear auditory or visual signals to ensure students respond appropriately.	Teacher typically uses clear auditory or visual signals to ensure students respond appropriately.	Teacher consistently uses clear auditory or visual signals to ensure students respond appropriately.
j) Teacher models skills/strategies during introduction of activity (e.g., shows students examples that demonstrate how to complete the academic skill/strategy, which all students can easily see, during teaching).	Teacher does not clearly demonstrate skills/strategies prior to student practice opportunities.	Teacher occasionally clearly demonstrates skills/strategies prior to student practice opportunities.	Teacher typically clearly demonstrates skills/strategies prior to student practice opportunities. ***Or*** no modeling is used but all students are successful with activities.	Teacher consistently demonstrates skills/strategies prior to student practice opportunities.
k) Teacher uses clear and consistent error corrections that demonstrates the correct response and has students practice the correct answer (e.g., use of corrective feedback procedures is evident and student(s) have the opportunity to respond correctly).	Teacher does not use corrective feedback procedures, including giving students an opportunity to practice the correct response.	Teacher occasionally uses corrective feedback procedures, including giving students an opportunity to practice the correct response.	Teacher typically uses corrective feedback procedures, including giving students an opportunity to practice the correct response ***or*** fewer than <u>three</u> errors occur during the entire lesson.	Teacher consistently uses corrective feedback procedures, including giving students an opportunity to practice the correct response.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
l) Teacher provides a range of systematic group or partner opportunities to respond (e.g., offers students practice by partner, choral and/or written responses).	Teacher does not provide opportunities for group or partner opportunities to respond.	Teacher provides some opportunities for group or partner opportunities to respond.	Teacher provides a range of systematic group or partner opportunities to respond.	Teacher regularly provides a range of systematic group or partner opportunities to respond.
m) Teacher presents individual turns systematically (e.g., students are given opportunities to respond individually but using a varied approach to keep students engaged, provides additional opportunities for students making regular errors).	Teacher does not present individual turns when appropriate.	Teacher occasionally presents individual turns when appropriate (round robin and turns are predictable).	Teacher presents individual turns when appropriate, purposely varied across students during some portions of the instruction. (All students are given opportunities to respond individually on a random basis.)	Teacher presents individual turns when appropriate purposely and strategically across students. (All students are given opportunities to respond individually on a random basis.) Individual turns are strategically incorporated throughout the instructional time.
n) Teacher systematically modulates lesson pacing/provides adequate think time (e.g., appropriate to learner performance).	Teacher makes no attempt to adjust pacing in response to student performance.	Teacher adjusts pacing/wait time occasionally in accordance with student responses.	Teacher typically anticipates and adjusts pacing/wait time between question and student response.	Teacher consistently anticipates and adjusts pacing/wait time between question and student response.
o) Teacher ensures students are firm on content prior to moving forward (e.g., holds students to a high criterion/mastery level of performance on each task, reteaches and retests as needed).	Teacher moves on before most students are firm on content.	Teacher moves on when some of the students are firm on the content or sometimes moves on when students are firm on content but other times moves on before students are firm on content.	Teacher typically ensures most students are firm on content before moving on to new material.	Teacher consistently moves on when most students are firm on the content or continues to practice when students are not firm on content. (if only one student persists in errors and the teacher moves on after attempting correction, this is ok)

If one activity goes particularly poorly, the ***teacher cannot receive a rating of 3* on the following items: familiarity with the lesson, clear and consistent wording, modeling, clear signals and correction procedures.

Student Response During Intervention

Group Student Behavior

Item	None or One 0 points <50%	Some 1 point >50%	Most 2 points >80%	All 3 points >95%
a) Students are familiar with group routines (e.g., students demonstrate they know procedures).	Students do not demonstrate knowledge of group routines.	Students occasionally demonstrate knowledge of group routines.	Most students typically demonstrate knowledge of group routines.	All students demonstrate knowledge of group routines consistently during the instruction.
b) Students are actively engaged with the lesson (e.g., students are listening, on task and responding).	Students are not actively engaged during the lesson.	Students are actively engaged during part of the lesson.	Most students are actively engaged for the majority of the lesson.	All students are actively engaged for the majority of the lesson.
c) Students follow teacher directions (e.g., students are listening and responding to teacher requests).	Students do not follow teacher's directions when asked.	Students occasionally follow teacher's directions when asked.	Most students typically follow teacher's directions when asked.	All students consistently follow all teacher's directions when asked.
d) Students are emotionally engaged with the teacher (e.g., students connect with teacher beyond schoolwork and are excited to be there).	Students don't appear to want to be in the group (e.g., students direct negative comments/behavior toward teacher, etc.).	Students seem complacent/compliant with the group (e.g., student "going through the motions" in group but not negative).	Most students appear to genuinely want to be in the group (e.g., students smile when joining the group, say hi to teacher, etc.).	All students appear to genuinely want to be in the group (e.g., students smile when joining the group, say hi to teacher, etc.).

Individual Student Response

Item	0 points <50%	1 point >50%	2 points >80%	3 points >95%
Emotional Engagement	Student appears to be disconnected from the teacher. Student responds to teacher attention with negative comments or behaviors.	Student appears to be somewhat connected with the teacher, but appears to be complacent with teacher attention. Student may not actively seek out teacher attention, but does not respond negatively to the teacher.	Student typically appears to be connected with the teacher and seems to seek interactions with teacher. Student smiles when joining group, appears happy to be there, seeks teacher attention, and appears to want to work with teacher.	Student consistently appears to be highly connected with the teacher and seems to seek interactions with teacher. Student smiles when joining group, appears happy to be there, seeks teacher attention, and appears to want to work with teacher.
Self-Regulated Behavior	Student demonstrates limited attention. Across the instructional observation, engagement is dependent upon significant teacher prompting. Consistently needs to be redirected to complete tasks.	Student demonstrates occasional attention to tasks (and may be able to maintain attention during one or certain type of tasks), but engagement is often dependent upon significant teacher prompting (e.g., at least 2 prompts in 1 task). Consistently needs to be redirected to complete tasks. After prompting, will comply.	Student demonstrates moderate engagement. Student is typically engaged but is sometimes dependent on teacher prompting (e.g., <2 within a task). Completes work/answers on signal, asks questions when appropriate. Appears to be trying hard. Sometimes volunteers to participate.	Student demonstrates consistent sustained attention. Able to stay engaged in lesson regardless of amount of teacher attention. Completes work/answers on signal, asks questions when appropriate. Appears to be trying hard. Student actively initiates and regularly volunteers to participate.

*Only code student individual behaviors if they are visible for the majority of the session (i.e., more than 50% of time).

Student Responsiveness Descriptors:

- Responsive: Student may or may not visibly demonstrate awareness of feedback, but attempts to incorporate feedback (i.e., accuracy improves, self-corrects) later in lesson.
- Non-responsive: Student may or may not demonstrate overt awareness of feedback, but demonstrates consistent error patterns across lesson.

•

Group ID: _____ Date of Video/Observation: _____ Observer Name: _____

Number of Minutes of Lesson:_____ Number of Students Observed: _____

Approximate time per activity type: Whole group:_____ Independent work: _____ Partner work:_____

Criteria for Level of Implementation Ratings (see developed **rubric** for each rating of implementation):

3 = Expert; 2 = Effective; 1 = Inconsistent; 0 = Element absent or not observed

Quality of Intervention Delivery If one activity goes particularly poorly, the ***teacher cannot receive a rating of 3*** on the following item: teacher familiarity of lesson, clear and consistent wording, modeling, clear signals and correction procedures.		
Item	***Level of Implementation***	***Comments***
a) Teacher is familiar with the lesson	0 1 2 3	
b) Instructional materials are organized	0 1 2 3	
c) Transitions from one activity to another are efficient and smooth (i.e., less than 2-3 minutes)	0 1 2 3	
d) Teacher expectations are clearly communicated and understood by students	0 1 2 3	
e) Teacher positively reinforces correct responses and behavior as appropriate (group and individual)	0 1 2 3	
f) Teacher appropriately responds to problem behavior (including off task)	0 1 2 3	
g) Teacher is responsive to the emotional needs of the students	0 1 2 3	
h) Teacher uses clear and consistent lesson wording	0 1 2 3	
i) Teacher uses clear auditory or visual signals	0 1 2 3	
j) Teacher models skills/strategies to introduce an activity	0 1 2 3	
k) Teacher uses clear and consistent error corrections that includes the correct response and has students practice the correct answer	0 1 2 3	
l) Teacher provides a range of systematic group or partner opportunities to respond	0 1 2 3	
m) Teacher presents individual turns systematically	0 1 2 3	
n) Teacher systematically modulates lesson pacing/provides adequate think time	0 1 2 3	
o) Teacher ensures students are firm on content prior to moving forward	0 1 2 3	
Overall Quality of Intervention Delivery Total	/45	

Overall Intervention Delivery

Overall effectiveness takes into consideration quality of delivery, understanding of the program, and student engagement and management.				
Ineffective	***Needs Improvement***	***Proficient***	***Effective***	***Highly Effective***
1	**3**	**5**	**7**	**9**

0 1 2 3 4 5 6 7 8 9 10

Student Response During Intervention

Group Student Behavior			
Item	***Level of Implementation***		***Comments***
a) Students are familiar with group routines	0 1 2 3		
b) Students are actively engaged with the lesson	0 1 2 3		
c) Students follow teacher directions	0 1 2 3		
d) Students are emotionally engaged with the teacher	0 1 2 3		
Overall Group Student Behavior	/12		

Individual Student Response (Record students from left to right from your perspective)			
Stud	**Emotional Engagement**	**Self-Regulated Behavior**	**Responsiveness**
S1	0 1 2 3	0 1 2 3	Responsive Non-Resp
S2	0 1 2 3	0 1 2 3	Responsive Non-Resp
S3	0 1 2 3	0 1 2 3	Responsive Non-Resp
S4	0 1 2 3	0 1 2 3	Responsive Non-Resp
S5	0 1 2 3	0 1 2 3	Responsive Non-Resp

**If student performance was unclear due to camera angle, indicate by placing an X over the student number. Only code student individual behaviors if they are visible for the majority of the session (i.e., more than 50% of time).

REFERENCES CITED

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: The Guilford Press.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., gitomer, D. H. & Pianta, R.C. Improving observational score quality: challenges in observer thinking. In T. Kane, K. Kerr, and R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project.* (pp. 415-443). Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Borman, G. D., Kimball, S. M., Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality : Do teachers with higher standards-ratings close student achievement gaps? *The Elementary School Journal*, *106*(1), 3–20.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.

Brooks, M. G., & Brooks, J. G. (1999). The courage to be constructivist. *The Constructivist Classroom*, *57*(3), 18–24.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*(10), 1069–1077. doi:10.1037//0003-066X.41.10.1069

Brophy, J. (1999). *Teaching* (Vol. 37). Geneva, Switzerland. doi:10.1016/S0167-8922(00)80004-8

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 328–375). New York, NY, US: Macmillan Publishing Company.

Brownell, M. T., Bishop, A. G., Gersten, R., Klingner, J., Penfield, R., Dimino, J., … Sindelar, P. T. (2009). The role of domain expertise in beginning special education teacher quality. *Exceptional Children*, *75*(4), 391–411.

Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, *43*(1), 61–85. doi:10.1016/j.jsp.2004.12.002

Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the complexity of instruction: A Study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading*, *15*(5), 409–439. doi:10.1080/10888438.2010.497521

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly*, *20*(3), 345–360. doi:10.1016/j.ecresq.2005.07.005

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2008). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*.

Causton-Theoharis, J. N., Doyle, M. B., Giangreco, M. F., & Vadasy, P. F. (2007). The "sous-chefs" of literacy instruction. *Teaching Exceptional Children*, *40*(1), 56–62.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. Cambridge, MA.

Chomat-mooney, L. I., Pianta, R. C., Hamre, B. K., Mashburn, A. J., Luckner, A. E., Grimm, K. J., … Downer, J. T. (2008). *A practical guide for conducting classroom observations: A summary of issues and evidence for researchers*. Charlottesville.

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed. *Psychological Assessment*, *6*(4), 284.

Colton, A. B., & Sparks-Langer, G. M. (1992). Restructuring student teaching experiences. In *Supervision in Transition* (pp. 155–168). Alexandria, VA: Association for Supervision and Curriculum Development.

Congdon, P. J., & MeQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*,*37*(2), 163-178.

Connor, C. M. (2013). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly*, *28*(4), 342–6. doi:10.1037/spq0000045

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., … Morrison, F. J. (2009). Individualizing student instruction precisely: effects of Child x Instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100. doi:10.1111/j.1467-8624.2008.01247.x

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, *79*(2), 135–144.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association of Supervision and Curriculum Development.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. Association of Supervision and Curriculum Development.

Danielson, C. (2013). *The Framework for Teaching evaluation instrument, 2013 edition: The newest rubric enhancing the links to the Common Core State Standards, with clarity of language for ease of use and scoring*.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Association of Supervision and Curriculum Development.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Retrieved from www.americanprogress.org

Denton, C. a., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities*, *39*(5), 447–466. doi:10.1177/00222194060390050601

Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching*. Holt, Rinehart & Winston.

Durlak, J. a, & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3-4), 327–50. doi:10.1007/s10464-008-9165-0

Engelmann, S., Arbogast, A., Bruner, E., Lou Davis, K., Engelmann, O., Hanner, S., & Al., E. (2002). *SRA Reading Mastery Plus*. DeSoto, TX: SRA/McGraw-Hill.

Evertson, C., & Harris, A. (1992). What we know about managing classrooms. *Educational Leadership*, *49*(7), 74–78.

Feng, L., Figlio, D. N., & Sass, T. R. (2010). *School accountability and teacher mobility*.

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage Publications Inc.

Fish, M. C., & Dane, E. (2000). The Classroom Systems Observation Scale : Development of an instrument to assess classrooms using a systems perspective. *Learning Environments Research*, *3*, 67–92.

TNTP (2013). *Fixing classroom observations: How common core will change the way we look at teaching*. Retrieved from http://tntp.org/assets/documents/TNTP_FixingClassroomObservations_2013.pdf

Fixsen, D. L., Blase, K., Metz, A., & Van Dyke, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*(2), 213–230.

Foorman, B. R., & Torgesen, J. (2001). Critical Elements of Classroom and Small-Group Instruction Promote Reading Success in All Children. *Learning Disabilities Research and Practice*, *16*(4), 203–212. doi:10.1111/0938-8982.00020

Forbes-Spear, C. (2014). *Examining the relationship between implementation and student outcomes: The application of an implementation measurement framework*. University of Oregon.

Foundation, B. and M. G. (2009). Measures of effective teaching project (MET).

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy , student achievement , and student awareness of learning. *American Educational Research Journal*, *21*(2), 449–460.

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Research and Evaluation : An International Journal on Theory and Practice*, *18*(7), 4–10.

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching : A review of criticisms. *Elementary School Journal*, *89*(3), 253–300.

Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. *Journal of Teacher Education*, *65*(5), 389–401. doi:10.1177/0022487114542519

Gay, L. R., Mills, G. E., & Airasian, P.W. (2009). *Educational research: Competencies for analysis and applications.* Upper Saddle River, NJ: Pearson Higher Education.

George, D., & Mallery, P. (2010). SPSS for windows step by step: A simple guide and reference. 18.0 update (11th ed.). Boston: Allyn & Bacon.

Gersten, R., Baker, S. K., Haager, D., & Graves, A. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners. *Remedial and Special Education*, *26*(4), 197–206.

Gersten, R., Fuchs, L., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education. *Exceptional Children*, *71*(2), 149–164.

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997). What we know about using research findings: Implications for improving Special Education practice. *Journal of Learning Disabilities*, *30*(5), 466–476.

Girolametto, L., & Weitzman, E. (2002). Responsiveness of child care providers in interactions with toddlers and preschoolers. *Language, Speech, and Hearing Services in Schools*, *33*(October), 268–281.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness : A research synthesis*. Retrieved from http://files.eric.ed.gov/fulltext/ED521228.pdf

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning*.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness* (pp. 1–12).

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best Practices in Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an Outcomes-Driven Model. In *Best practices in school psychology IV (Vol. 1, Vol. 2).* (pp. 699–720). Washington, DC, US: National Association of School Psychologists.

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Laimon, L., Smith, S., & Flindt, N. (2004). *Technical adequacy and decision making utility of DIBELS (No. 7)*.

Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology Review*, *31*(3), 328–349.

Gudmundsdottir, S. (1997). Introduction to the theme issue of "narrative perspectives on research on teaching and teacher education." *Teaching and Teacher Education*, *13*(I), 1–3.

Hagan-Burke, S., Coyne, M. D., Kwok, O.-M., Simmons, D. C., Kim, M., Simmons, L. E., … McSparran Ruby, M. (2013). The effects and interactions of student, teacher, and setting variables on reading outcomes for kindergarteners receiving supplemental reading intervention. *Journal of Learning Disabilities*, *46*(3), 260–77. doi:10.1177/0022219411420571

Hall, T., Vue, G., Strangman, N., & Meyer, A. (2014). Differentiated instruction and implications for UDL implementation. *Effective Classroom Practices Report*. Retrieved November 07, 2014, from http://aim.cast.org/learn/historyarchive/backgroundpapers#.VF0K8jTF9Og

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol*, *8*(1), 23–34.

Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). *Classroom Assessment Scoring System (CLASS) implementation guide: Measuring and improving classroom interactions in early childhood settings*. Charlottesville, VA. Retrieved from teachstone.com/wp-content/uploads/.../CLASSImplementationGuide.pdf

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*(5), 949–967.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms* (pp. 1–35). Retrieved from http://www.researchgate.net/profile/Jason_Downer/publication/237728991_Building_a_Science_of_Classrooms_Application_of_the_CLASS_Framework_in_over_4000_U.S._Early_Childhood_and_Elementary_Classrooms/links/0046352cc1bf3e4168000000.pdf

Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?*. Washington, D.C.

Hanushek, E. A., & Rivkin, S. G. (2010). *Using value-added measures of teacher quality* (pp. 1–6). doi:10.1037/e722242011-001

Hanushek, E. A., Rivkin, S. G., The, S., Economic, A., The, P. O. F., & Rivkin, G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers & Proceedings*, *100*(2), 267–271.

Harms, T., & Clifford, R. (1980). *Early childhood environment rating scale*. New York: Teachers College Press.

Harms, T., Clifford, R., & Cryer, D. (1998). *Early Childhood Environmental Rating Scale-Revised*. New York: Teachers College Press.

Harn, B. A., Forbes-Spear, C., Fritz, R., & Berg, T. (2011). *Quality of Intervention Delivery and Receipt (QIDR) observation tool*. Eugene, OR.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*(2), 181–193.

Heneman, H., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay*. CPRE Policy Briefs. Retrieved from http://repository.upenn.edu/cpre_policybriefs/33

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems.*Harvard educational review*, *83*(2), 371-384.

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Using student growth to evaluate educators of students with disabilities : Issues, challenges, and next steps*.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 207–219.

Jackson, A. W., & Davis, G. A. (2000). *Turning points 2000: Education adolescents in the 21st century*. Williston, VT 05495-0020: Teachers College Press.

Jerald, C. (2012). *Ensuring accurate feedback from observations: Perspectives on practice.* Retrieved from https://docs.gatesfoundation.org/documents/ensuring-accuracy-wp.pdf

Jensen, E. (1998). *Teaching with the brain in mind*. Alexandria, VA: Association for Supervision and Curriculum Development.

Joe, J. N., McClellan, C. A., and Holtzman, S. L. Reliability and the length and focus of classroom observations (2016). In T. Kane, K. Kerr, and R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project.* (pp. 415-443). Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/11/Designing-Teacher-Evaluation-Systems_freePDF.pdf

Joe, J., Tocci, C., Holtzman, S., & Williams, J. (2013). *Foundations of Observation*. Princeton, NJ.

Johnson, E. S., & Semmelroth, C. L. (2012). Examining interrater agreement analyses of a pilot special education observation tool. *Journal of Special Education Apprenticeship*, *1*(2).

Johnson, E., & Semmelroth, C. L. (2013). Special Education Teacher Evaluation: Why It Matters, What Makes It Challenging, and How to Address These Challenges. *Assessment for Effective Intervention*, *39*(2), 71–82. doi:10.1177/1534508413513315

Jones, N. D., & Brownell, M. T. (2013). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention*, *39*(2), 112–124. doi:10.1177/1534508413514103

Justice, L. M. (2006). Evidence-based practice, response to intervention, and the prevention of reading difficulties. *Language, Speech, and Hearing Services in Schools*, *37*(October), 284–298.

Justice, L. M., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of Language and Literacy Instruction in Preschool Classrooms Serving At-Risk Pupils. *Early Childhood Research Quarterly*, *23*(1), 51–68. doi:10.1016/j.ecresq.2007.09.004

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/12/MET_Reliability-of-Classroom-Observations_Research-Paper.pdf

Jones, R. R., Reid, J. B., & Patterson, G. R. (1975). Naturalistic observation in clinical assessment. *Advances in psychological assessment, 3,* 42-95.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching. *Measures of Effective Teaching Project: Bill and Melinda Gates Foundation*.

Kane, T. J., Staiger, D. O., & McCaffrey, D. F. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data*. Cambridge, MA.

Knight, J. (2007). *Instructional Coaching*. Thousand Oaks, CA.

Kretlow, a. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *33*(4), 279–299. doi:10.1177/0888406410371643

La Paro, K., Pianta, R. C., & Stuhlman, M. (2004). Classroom assessment scoring system (CLASS): Findings from the pre-k year. *The Elementary School Journal*, (104), 409–426.

National Center for Early Development & Learning (1997). *Classroom observation system-kindergarten*. Charlottesville: University of Virginia.

Macmillan, C. J. B., & Garrison, J. (1984). Using the “new philosophy of science” in criticizing current research traditions in education". *Educational Researcher, 13*(10), 15–21.

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in schools*. Association of Supervision and Curriculum Development.

Mashburn, A. J., Pianta, R. C., Barbarin, O. A., Bryant, D., Hamre, K., Downer, J. T., … Howes, C. (2008). Measures of classroom quality in prekindergarten and children’s development of academic , language , and social skills. *Child Development, 79*(3), 732–749.

Maxwell, K. L., McWilliam, R. A., Hemmeter, M. L., Ault, M. J., & Schuster, J. W. (2001). Predictors of developmentally appropriate classroom practices in kindergarten through third grade. *Early Childhood Research Quarterly, 16*, 431–452.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies.*Psychological methods*, *9*(2), 147.

McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training and certification: Lessons learned from teh Measures of Effective Teaching project*. San Francisco, CA.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, (1), 30.

McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems*.

Medley, D. M. (1979). The effectiveness of teachers. In *Research on teaching: Concepts, findings, and implications* (pp. 11–27).

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement : Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53.

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology*. Sage.

Mowbray, C. T., Holter, M. C., Teauge, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340. doi:10.1177/109821400302400303

Myers, D. M., Simonsen, B., & Sugai, G. (2011). Increasing teachers' use of praise with a response-to-intervention approach. *Education and treatment of children*, *34*(1), 35-59.

National Council on Teacher Quality *State of the states 2012 : Teacher effectiveness policies*. (2012). Washington, DC. Retrieved from http://www.nctq.org/dmsView/State_of_the_States_2012_Teacher_Effectiveness_Policies_NCTQ_Report

National Institute of Child Health and Human Development Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development*, *71*(4), 960–980.

NICHD Early Child Care Research, & Duncan, G. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, (74), 1454–1475.

NICHD Early Child Care Research Network. (1996). Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly*, *11*(3), 269–306. doi:10.1016/S0885-2006(96)90009-5

NICHD Early Child Care Research Network. (2002a). *Classroom observation system-first grade*. Charlottesville, VA: University of Virginia.

NICHD Early Child Care Research Network. (2002b). Early child care and children ' s development prior to school entry : Results from the nICHD study of early child care. *American Educational Research Journal*, *39*(1), 133–164.

NICHD Early Child Care Research Network. (2004). *Classroom observation system-fifth grade*. Charlottesville, VA: University of Virginia.

Odom, S. L. (2008). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education*, *29*(1), 53–61. doi:10.1177/0271121408329171

Officers, C. of C. S. S. (2011). *InTASC model core teaching standards: A resource for state dialogue*. Washington, DC.

Pianta, R. C. (2003). *Standardized classroom observations from pre-k to third grade: A mechanism for improving quality classroom experiences during the P-3 years*. Retrieved from http://fcd-us.org/sites/default/files/StandardizedClassroomObservations.pdf

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. J. (2007). Opportunities to learn in America's elementary classrooms. *Science*, *315*, 1795–1796.

Pianta, R. C., Belsky, J., Houts, R., Morrison, F., & National Institute of Child Health and Human Development Early Child Care Research Network. (2007). Opportunities to learn in America's elementary classrooms. *Science*, *315*, 9–10.

Pianta, R. C., Cox, M. J., Taylor, L., & Early, D. (2013). Kindergarten teachers ' practices related to the transition to school : Results of a national survey. *The Elementary School Journal*, *100*(1), 71–86.

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom Assessment Scoring System*.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*, 431–451.

Pianta, R. C., Paro, K. M., Payne, C., Cox, M. J., Bradley, R., Pianta, R. C., … Payne, C. (2002). The relation of kindergarten classroom environment to teacher , family , and school characteristics and child outcomes. *The Elementary School Journal*, *102*(3), 225–238.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs , Classrooms , and teachers : Do they predict observed classroom quality and child-teacher interactions. *Applied Developmental Science*, *9*(3), 144–159.

Pratt, A., & Logan, J. (2014). Improving language-focused comprehension instruction in primary-grade classrooms: Impacts of the Let's Know! experimental curriculum. *Educational Psychology Review*, *July*.

Pressley, M., Roehrig, A. D., Raphael, L., Dolezal, S., Bohn, C. M., Mohan, L., & Hogan, K. (2003). Teaching processes in elementary and secondary education. In *Handbook of psychology* (Vol. 1). doi:10.1037/005272

Raudenbush, S., Bryk, A., & Congdon, R. (2013). HLM 7.01 for Windows [Hierarchical linear and nonlinear modeling software]. *Skokie, IL: Scientific Software International*.

Reid, J. B., Skindrud, K. D., Taplin, P. S., & Jones, R. R. (1973, August). The role of complexity in the collection and evaluation of observation data. In*meeting of the American Psychological Association, Montreal, Canada*.

Repp, A. C., Nieminen, G. S., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children*, *55*(1), 29-36.

Rimm-Kaufman, S. E., Early, D. M., Cox, M. J., Saluja, G., Pianta, R. C., Bradley, R. H., & Payne, C. (2002). Early behavioral attributes and teachers' sensitivity as predictors of competent behavior in the kindergarten classroom. *Journal of Applied Developmental Psychology*, *23*(4), 451–470. doi:10.1016/S0193-3973(02)00128-4

Rosenshine, B. (1971). *Teaching behaviours and student achievement*.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M. C. Witrock (Ed.), *Handbook on research and teaching* (3rd ed., pp. 376–390). New York, NY: Macmillan.

Ross, J., & Regan, E. (1993). Sharing professional experience: Its impact on professional development. *Teaching and Teacher Education*, *9*(1), 91–106.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliott, K. (2002). *The effective provision of pre-school education (EPPE) project: Measuring the impact of pre-school on chidren's cognitive process over the pre-school period* (Vol. 44). London, UK.

Saunders, K. J. (2011). Designing instructional programming for early reading skills. In W. Fisher & C. Piazza (Eds.), *Handbook of Applied Behavior Analysis* (pp. 92–109). New York, NY, US: Guilford Press.

Schmoker, M. J. (1999). *Results: The key to continuous school improvement*. Association of Supervision and Curriculum Development.

Semmelroth, C. L., & Johnson, E. (2013). Measuring Rater Reliability on a Special Education Observation Tool. *Assessment for Effective Intervention*, *39*(3), 131–145. doi:10.1177/1534508413511488

Shaywitz, S. (2008). *Overcoming dyslexia: A new and complete science-based program for reading problems at any level*. New York: Random House.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational review1*, *57*(1), 1–23.

Simmons, D. C., Coyne, M. D., Hagan-Burke, S., Kwok, O.-M., Simmons, L. E., Johnson, C., … Crevecoeur, Y. C. (2011). Effects of supplemental reading interventions in authentic contexts: A comparison of kindergarteners' response. *Exceptional Children*, *77*(2), 207–228.

Simmons, D. C., & Kame'enui, E. J. (2003). *Scott Foresman: Early Reading Intervention*. Glenview, IL: Scott Foresman.

Skowron, J. (2001). *Powerful lesson planning models*. Arlington Heights, IL: Skylight Publishing.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. doi:10.1002/pits.20113

Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Preventing School Failure: Alternative Education for Children and Youth*, *52*(2), 48–58. doi:10.3200/PSFL.52.2.48-58

Strong, M., Gargani, J., & Hacifazlioglu, O. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education*, *62*(4), 367–382. doi:10.1177/0022487110390221

Stronge, J. H. (2005). *Evaluating teaching: A guide to current thinking and best practice*. Corwin Press.

Swanson, H. L. (1999). Reading Research for Students with LD: A Meta-Analysis of Intervention Outcomes. *Journal of Learning Disabilities, 32*(6), 504–532. doi:10.1177/002221949903200605

Sykes, G., & Bird, T. (1992). Teacher education and the case idea. *Review of Research in Education*, *18*, 457–521.

Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics, 6th International edition (cover) edn.

Tomlinson, C. A. (1999). Mapping a Route Toward Differentiated Instruction. *Educational Leadership*, *57*(1), 12–16.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, *40*(1), 7–26. doi:10.1016/S0022-4405(01)00092-9

Torgesen, J. K., Wagner, R. K., Rashotte, C. a., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, *91*(4), 579–593. doi:10.1037//0022-0663.91.4.579

Waxman, H. C., Huang, S.-Y. L., Anderson, L., & Weinstein, T. (1997). Classroom process differences in inner-city elementary schools. *The Journal of Educational Research*, *91*(1), 49–59. doi:10.1080/00220679709597520

Wiggins, G., & McTighe, J. (1998). *Understanding by Design* (pp. 1–34). Association for Supervision and Curriculum Development.

Wolf, M. (2007). *Proust and the squid : the story and science of the reading brain / Maryanne Wolf ; illustrations by Catherine Stoodley*. (C. J. Stoodley, Ed.). New York, NY : HarperCollins. Retrieved from http://www.loc.gov/catdir/toc/fy0803/2008297333.html

Woodcock, R., & Johnson, M. (1989). *Woodcock-Johson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W. (1987). *Woodcock reading mastery tests, revised*. Circle Pines, MN: American Guidance Service.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*(2), 156–167. doi:10.1037//0022-006X.49.2.156

Yopp, H. K., & Yopp, R. H. (2000). Supporting Phonemic Awareness Development in the Classroom. *The Reading Teacher*, *54*(2), 130–143 CR – Copyright © 2000 International . Retrieved from http://www.jstor.org/stable/20204888

Zigmond, N., & Kloo, A. (2011). General and special education are (and should be) different. In Kauffman, J. M. *Handbook of special education* (pp. 160–172).