EXAMINING THE RELATIONSHIP BETWEEN IMPLEMENTATION AND STUDENT OUTCOMES: THE APPLICATION OF AN IMPLEMENTATION MEASUREMENT FRAMEWORK

by

CAITLIN FORBES SPEAR

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

December 2014

THESIS APPROVAL PAGE

Student: Caitlin Forbes Spear

Title: Examining the Relationship Between Implementation and Student Outcomes: The Application of an Implementation Measurement Framework

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:

Elizabeth Harn	Chairperson
Robert Horner	Core Member
Christopher Murray	Core Member
Michael Stoolmiller	Core Member
Gina Biancarosa	Institutional Representative

and

J. Andrew Berglund	Dean of the Graduate School

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

Degree awarded December 2014

DISSERTATION ABSTRACT

Caitlin Forbes Spear

Doctor of Philosophy

Department of Special Education and Clinical Sciences

December 2014

Title: Examining the Relationship Between Implementation and Student Outcomes: The Application of an Implementation Measurement Framework

The current study evaluated the implementation of evidence-based reading interventions using a multifaceted implementation measurement approach. Multilevel modeling was used to examine how three direct measures of implementation related to each other and to student academic outcomes and to examine patterns of implementation across time. Eight instructional groups were video taped weekly for nine weeks, and pre- and post-test assessments were given to 31 at-risk kindergartners from two schools using established evidence-based practices. Each implementation measure represented a different measurement approach (i.e., discrete behavioral measurement, global ratings) and focused on different aspects of implementation (e.g., structural, process, or multicomponent elements). Overall, results of this analysis indicated that (a) the implementation tools were highly correlated with each other, (b) only the multicomponent tool independently accounted for group differences, (c) together the multicomponent and process-oriented measures appear to account for additional variance in group differences, and (d) there were no significant trends in implementation across time as measured by any of the tools, however there were significant differences in trends over time between groups when using the structural measure. Implications for research and practice are discussed,

including the importance of taking a multifaceted approach to measuring implementation and aligning implementation measures with program theory.

CURRICULUM VITAE

NAME OF AUTHOR: Caitlin Forbes Spear

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
College of Staten Island, New York, NY
Wesleyan University, Middletown, CT

DEGREES AWARDED:

Doctor of Philosophy, Special Education, 2014, University of Oregon
Master of Science in Education, Special and General Education, 2007, College of Staten Island
Bachelor of Arts, Psychology and French Studies, 2003, Wesleyan University

AREAS OF SPECIAL INTEREST:

High Incidence Disabilities: Emotional and Behavioral Disorders
Integration of Academic and Behavioral Supports
Culturally Responsive Pedagogy
Teacher Education

PROFESSIONAL EXPERIENCE:

Postdoctoral Researcher, The Ohio State University, Columbus, Ohio
2014 - Present

Instructor, Special Education, University of Oregon, Eugene, Oregon
2011 – 2013

Continuing Education Online Course Facilitator, New York Teachers' Network
New York, New York 2009

Special Education Teacher, NYC Department of Education, Brooklyn, New York
2007 – 2009

General Education Teacher, NYC Department of Education, Brooklyn, New York
2006

Special Education Teacher, NYC Department of Education, Brooklyn, New York
2004 – 2006

Assistant Classroom Teacher, Shady Lane Preschool, Pittsburgh, PA
2003-2004

GRANTS, AWARDS, AND HONORS:

Clare Wilkins Chamberlin Memorial Research Award, Doctoral Research
University of Oregon, 2014

Jane Wiley Scholarship, College of Education Scholarship
University of Oregon, 2013

Florence Wolfard Scholarship, College of Education Scholarship
University of Oregon, 2013

Thompson Family Scholarship, College of Education Scholarship
University of Oregon, 2012

Silvy Kraus Presidential Fellowship, College of Education Scholarship
University of Oregon, 2012

PUBLICATIONS:

Spear, C. F., Strickland-Cohen, M. K., Romer, N., & Albin, R. (2013). An examination of social validity within single-case research with students with emotional and behavioral disorders. *Remedial and Special Education, 34,* 357-370. doi: 10.1177/0741932513490809

ACKNOWLEDGMENTS

My success in this endeavor would not have been possible without the support and unique contributions of my committee members: Dr. Elizabeth Harn, Dr. Gina Biancarosa, Dr. Chris Murray, Dr. Robert Horner, and Dr. Mike Stoolmiller. In particular I want to acknowledge my advisor Beth Harn, whose guidance, feedback, dedication, time and energy were endless. Thank you for your patience and support through all the many turns I took, and pushing me to think like a researcher. I also want to thank Ronda Fritz, Kate Ascetta, Abra Cooper, Jerin Kim, Michelle Massar, Manuel Monzalve, Aaron Mowery, Kara Rawlings, and Alison Wong for their dedication to data support for this project, and Dr. Deni Basaraba, Ronda Fritz, and Tricia Berg for their continual support and feedback.

Thank you also to Dr. Laurie Gutman Kahn, Dr. Jaime Lee, Dr. Yen Pham, and Dr. Kathleen Strickland-Cohen for their friendship, perspective, laughter, and support. I am so lucky to have completed this journey with each and every one of you. Finally, I want to thank my former students and colleagues at PS 10 and PS 396@304 in Brooklyn, New York for showing me each and every day that instruction matters, and inspiring me to ask bigger questions.

Dedicated to my daughter Ida, who puts it all in perspective; and to my partner Chris, for being there every step of the way.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Implementation

Implementation is a jack of many trades; difficult to define, challenging to measure, yet critically important to improving the delivery of evidence-based practices, particularly in the field of education. It is a term that has been used in various literature bases (e.g., mental health, K-12 education, early childhood) with multiple definitions, widely varying scope, and measurement considerations. Across these sources, fidelity of implementation (e.g., treatment fidelity, treatment integrity) is the most commonly accepted interpretation. While this term has a number of definitions, it is most commonly used as a synonym for adherence, or a measure of the degree to which a program or intervention is delivered as intended by program or researcher design (e.g., Harn, Parisi, & Stoolmiller, 2013; Mowbray, Holter, Teague, & Bybee, 2003; Odom et al., 2010; O'Donnell, 2008). From a researcher perspective, this focus on adherence is critical for documenting the internal validity of the study to measure whether the obtained effect is a function on the intervention being implemented as expected (Mowbray et al., 2003; O'Donnell, 2008). Other fields, such as program evaluation, or more recently implementation science, conceive of implementation as a larger process that examines service delivery (i.e., use of the program/intervention) over time and across systems (i.e., scale-up), with a focus on the interactions between the multiple stakeholders (e.g., teachers, principals) whose relationships impact participant outcomes (Durlak & DuPre, 2008; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Harn et al., 2013; Mowbray et al., 2003).

Whether viewed from the fine (i.e., fidelity/adherence) or broad (i.e., systems/scale-up) perspective, the rationale for studying implementation – to determine what actually occurs during program delivery – has important implications for researchers, practitioners, and policy makers alike (Odom, Hansen et al., 2010). Researchers have begun the important work of documenting the efficacy of many of these programs to promote their use in schools, however much is as yet unknown about the complex variables that impact efficient and sustainable implementation of these programs in real world settings (Fixsen et al., 2005; Odom, 2008; Odom, Hansen et al., 2010). Despite these unknowns, many of these programs are described as evidence-based when the only aspect of implementation that has actually been assessed is adherence to program protocol within the constraints of a research study. This approach highlights a problematic assumption that practices will transfer across contexts, or that adherence to program, which is clearly an important variable when determining program efficacy, is related, or synonymous, to sustainability (Harn, et al., 2013; O'Donnell, 2008).

Ultimately, the complex act of implementation is as important to the process of disseminating and sustaining the use of evidence-based practices (EBPs) as the development of programs themselves, yet very few standards of evidence for assessing and measuring that process are widely in use (Flay et al., 2005). The lack of agreement on how to define implementation, the erroneous use of the term as a synonym for fidelity, and the problematic assumption that measures of adherence are adequate to capture the full range of variables that impact implementation pose significant challenges to the field as it works to support schools in selecting and sustaining the use of EBPs. While the field has yet to develop this common definition, there is growing agreement that a

comprehensive model of implementation must include two core mechanisms that are discussed next: delivery (what to measure) and measurement (when and how to measure) (Odom, Hansen et al., 2010).

Delivery. Delivery is a key element of any definition of implementation. Whether focused on an intervention, an innovation, or an approach, implementation involves the delivery of services. For the purposes of clarity, from this point on these services will be referred to generally as a program. While this is a clear starting point, this notion of delivery is complex, and typical approaches to defining and measuring implementation often overlook the nuances involved in program delivery, which include both the content and context.

Content – What to measure? The first element of delivery that needs to be considered is content, or the "what" of the program. In the field of education, this typically involves curricular content, or instruction on a sequence of information designed to ensure that students learn specific concepts, behaviors, skills, or strategies (Howell & Nolet, 2000). Typical views of implementation assess the degree to which a packaged set of tasks, designed by researchers or curriculum developers, are delivered as intended. While these definitions are useful, and fairly widely accepted, they are really only appropriate for determining whether program delivery adhered to a set protocol, and the amount of the program that was delivered. These types of implementation measures should more accurately be thought of as the first step in determining whether a program is efficacious. Additional steps are needed to determine the active ingredients of the program that are the underlying mechanisms of change, rather than simply examining implementation of the full package (Durlak, 2010).

This is true for programs that are both curriculum dependent (e.g., *Reading* Mastery), as well as those that are curriculum independent, and not necessarily specified by procedure (e.g., explicit instruction, positive behavior support). These types of programs still contain active ingredients or critical components that must be delivered and accounted for, however assessing adherence to a checklist of procedures may overlook important delivery features for these types of programs, such as variables related to how well programs are delivered or how students attended to different types of instruction (Harn & Parisi, 2013; Harn et al., 2013; Odom et al., 2010). Recent work in the field of implementation science has led to the development of definitions that encompass these broader views of delivery. For example Fixsen and colleagues define implementation as "a specified set of activities designed to put into practice an activity or program of known dimensions" (2005, p. 5). This is an important step toward developing a comprehensive model of implementation, however, it does not address another critical aspect to implementation delivery: contextual variables.

Context. Program delivery is not simply a function of delivering content, but is also impacted by contextual factors (e.g., who delivers it, who receives it, structural and environmental factors) (Harn et al., 2013). A comprehensive definition of implementation must therefore also account for the contextual nature of delivery (Odom, Hansen, et al. 2010). When delivering instruction in school settings, contextual variables (e.g., district, school, classroom, teacher, student) can impact implementation and outcomes and need to be considered (Flay, et al., 2005; Harn et al., 2013; Odom, Hansen et al., 2010). Durlak and DuPre (2008) extend Fixsen et al.'s (2005) definition to encompass the context as well as the program by defining implementation as "what a program consists

of when it is delivered in a particular setting" (p. 239). Odom, Hansen and colleagues (2010) take this one step further, adding, "the program delivered to *and experienced* by participants" (p. 417, emphasis added). Each of these definitions allows for a flexible view of contextualized delivery, however these broad views of implementation fail to specify how these broader contextual variables may impact measurement or their impact on outcomes.

Measurement—When and how to measure? While it is often discussed as such, implementation is not simple program delivery. It is also the complex act of *evaluating* that delivery – a process that involves not only deciding the purpose of that evaluation (e.g., Are we evaluating initial implementation or supporting schools in scaling-up a program?), but also where the school is in the process, and questions about when and how to measure that delivery. Understanding these factors should impact what and how implementation is measured (Flay et al. 2005). As such, the second element of implementation that needs to be considered is how it will be measured. Two dimensions of measurement are particularly important when considering implementation: time and specific measurement approach.

Time—When to measure implementation. There are several time-related issues to consider when measuring implementation. Typical approaches to assessing implementation often report one overall score that is either representative of just one assessment point, or is averaged across an entire study. There is growing recognition that implementation is not stable, but rather dynamic, and changes over time (e.g., within a school day, over the course of a school year, from beginning to end of an intervention)

(Chomat-Mooney et al., 2008; Harn et al., 2013; Odom et al., 2010; Zvoch, 2009). This has important implications for the way implementation should be measured and reported.

Implementation is a complex process, not a singular act of establishing program delivery (Fixsen et al., 2005; Odom, Hansen et al., 2010). Fixsen and colleagues outline six stages of implementation that move a program from initial development to sustained use: (a) exploration, (b) installation, (c) initial implementation, (d) full operation, (e) innovation, and (f) sustainability. These stages, which will be described fully in the next chapter, have unique and distinct components involved throughout the process of implementation.

Clearly, time is an important variable in a process this complex. Fixsen et al. (2005) estimate that the entire implementation process can take two to four years for a single EBP, but there is limited research investigating this process, with the majority of efforts focusing on initial implementation. These findings may highlight issues with a specific program, but it's also quite possible that time within implementation stage is a factor here, and that programs may have different effects on student outcomes at different stages of implementation (Durlak & DuPre, 2008; Flay et al., 2005).

Approaches—How to measure implementation. The second measurement element that needs to be considered in a comprehensive view of implementation is the actual measurement process itself, namely the types of measures and the measurement approaches that are used to assess implementation. While discussed in more depth in the next chapter, two recent reviews of implementation literature in the mental health and education fields indicate that there are four types of measures that are commonly used to assess implementation: (a) direct observation, (b) self-report, (c) interviews, and (d)

assessment of archival records (Durlak & DuPre, 2008; O'Donnell, 2008; Odom, Hansen et al., 2010). Of these four approaches, only direct observation assesses implementation in *action*, enabling observers to gather data that considers the full complexities of program delivery (i.e., content and context), which may support long-term sustainability (Flay et al, 2005).

The information that is obtained during implementation measurement depends on the theoretical framework, perspective, and measurement approach that is used to obtain the information. Different methodologies emphasize different aspects of implementation (e.g., rich descriptions via qualitative methods, frequency counts of specifically defined behaviors using quantitative methods, etc.), and measurement tools can therefore be tailored to highlight and capture different aspects of implementation. Different measurement approaches can also be used to assess various time frames, from a broad perspective on global features of classrooms across time, to discrete assessments of classroom processes in momentary time samples. Each of these elements has important implications for the development of an implementation measurement system. For example, issues such as construct validity, reliability, and technical aspects of observational approaches (e.g., rater effects on observations, stability of indicators) are addressed in different ways by various approaches (Chomat-Mooney et al., 2008). These measurement questions must be carefully considered in any comprehensive view of implementation. It is only with a clearly defined construct, and accurate, reliable and valid measurement that the field can truly begin to assess, understand, and impact implementation on a large scale (Odom, Hansen et al., 2010).

Implementation as a multifaceted construct. Given the complex variables involved in delivering and measuring implementation, it is necessary to recognize the multifaceted nature of this construct when developing a comprehensive model. Several researchers have begun work in this vein, drawing from the work of Dane and Schneider (1998) and Durlak and DuPre (2008) to describe eight distinct aspects of implementation that can be measured individually: (a) adherence, (b) dosage, (c) quality, (d) participant response, (e) program differentiation, (f) monitoring comparison conditions, (g) program reach, and (h) adaptation. These subcomponents of implementation, discussed more in the next chapter, can be useful in considering how the content and context of program delivery, and measurement timing and approaches can impact implementation assessment (Odom, Hansen et al., 2010). This multifaceted view of implementation may hold important implications for the field of education, given the limited understandings of which aspects of implementation are most related to student outcomes and sustainability (Durlak, 2010; Webster-Stratton, Reinke, Herman, & Newcomer, 2011).

This limited understanding stems from the assumption that implementation measures of adherence in an efficacy study are related to the transfer of effective use of EBPs in real word settings. This assumption is based on the supposition that a program has been systematically studied to identify the active core ingredients that *must* be included to achieve the determined effect, rather than only studied as an integrated package; however very few programs identified as evidence-based have completed this complex task (Durlak, 2010; Harn & Parisi, 2013; Harn et al., 2013). This assumption also presumes that these adherence measures assist in long-term sustainability of program use, though both Durlak (2010) and Fixsen et al. (2005) note the problematic nature of

this assumption, and discuss its role in maintaining the research to practice gap. Using a comprehensive model that conceives of implementation as a complex, multifaceted construct may provide important insight into how to successfully transfer significant research outcomes into sustained, effective school-based delivery (Flay et al., 2013; Harn et al., 2013).

Toward a Comprehensive Model

As Durlak and DuPre note, "Science cannot study what it cannot measure accurately and cannot measure what it does not define" (2008, p. 342). Implementation scientists like Fixsen and colleagues (2005), Durlak and DuPre (2008), and Odom, Hansen et al. (2010) have laid the groundwork for this important process of defining and accurately measuring implementation, yet there is much more work to be done. While the field has begun to discuss implementation as a multifaceted construct with elements of delivery and measurement, very few researchers are combining all of these elements in their examinations of implementation, leaving a large gap in the knowledge base as to which implementation measures are appropriate for which contexts, and how different tools relate to long-term outcomes (Flay et al., 2005). Figure 1 offers a useful platform for beginning this type of work, as it combines models of implementation science, measurement considerations, and multicomponent views of implementation to provide an overarching framework of the delivery and measurement components that must be addressed in a comprehensive view of implementation.

Statement of Purpose

With this conceptual framework in mind, this study proposed to examine the implementation of established evidenced-based early literacy programs being fully

implemented in school settings using different evaluation and observational approaches. By examining the implementation of an established EBP, this study focused on how well teachers delivered the programs and its impact on outcomes. In terms of measurement, the three implementation approaches all involved the direct observation of the program, however they varied in terms of their measurement approach and their emphasis on the different dimensions of implementation. These different implementation tools examined four essential aspects of implementation: (a) adherence to program theory, (b) dosage, (c) quality, and (d) participant responsiveness. The current study also examined how these aspects relate to student outcomes during the full operation stage of implementation. This stage was selected as the processes that make program delivery more or less effective are expected to be more salient during the full operation stage (Fixsen et al., 2005).

Given these specifications about delivery and measurement, the purpose of the current study was to examine the extent to which instructional implementation during small group reading interventions, as measured by three different implementation tools (OTR, CLASS, QIDR), is associated with the academic outcomes of at-risk kindergarten students. This was addressed using two level hierarchical linear modeling, with students nested in groups. Patterns of implementation across time were also examined using a two level growth model, looking at time across group. The next chapter provides a synthesis of the implementation literature base in which to situate the current study. This review of the literature is followed by a chapter that details the methods of the current study, which outlines both the intervention and implementation data collection procedures that were used in the present study, as well as the analysis approach. The fourth chapter provides an

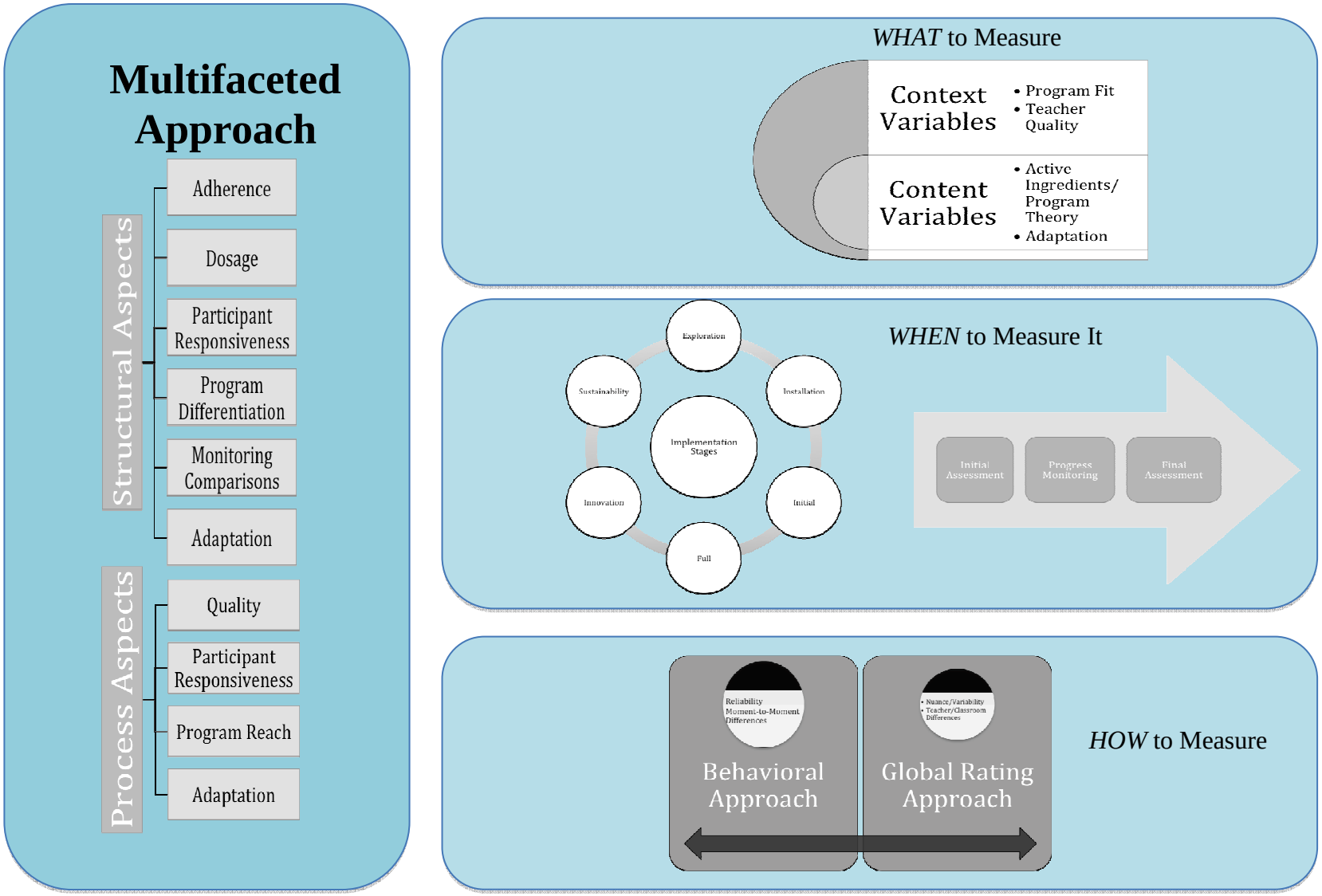

Figure 1. Overview of relationships between key delivery and measurement components within a multifaceted implementation framework. Adapted from Chomat-Mooney et al., 2008; Durlak, 2010; Durlak & DuPre, 2008; Fixsen et al., 2005; Mowbray et al., 2003; O'Donnell, 2008; Zvoch, 2009.

overview of the results of all analyses, while the final chapter concludes with a discussion of the findings and their implications for research and practice.

CHAPTER II

LITERATURE REVIEW

> "[Research suggests] that the what, who, when, and how of implementation need more careful inquiry. In other words, we need more clarity about which aspects of implementation are most important for different outcomes, how to assess each aspect most accurately, who should provide the necessary data, when these assessments should be done, and what ecological factors should be evaluated. Then decisions must be made about the best analytic approach to apply when studying the relationship between implementation and different outcomes." (Durlak, 2010; p. 353)

Implementation Matters

In today's era of educational accountability, the heart of school reform efforts are aimed at improving student learning, with a focus on producing academic gains and long-term outcomes. Student learning is a complex ecological process, impacted by multiple social systems: family, community, school, policy – all of which exist within distinct, yet interconnected cultural contexts (Guskey, 2000). Despite this complexity, educational institutions are being held accountable for producing similar outcomes across multiple educational systems – state, district, individual school, classroom, and teacher – and multiple factors influence student learning at each of these levels. While all of these systems interact, and have a huge impact on student outcomes, the majority of student learning takes place in the classroom. Within each classroom, teacher instruction is the main mechanism for impacting that learning process (e.g., Cohen & Ball, 1999; Fishman, Marx, Best, & Tal, 2003; Garet, Porter, Desimone, Birman, & Yoon, 2001; Pianta & Hamre, 2009).

Instruction varies widely, but the assumption in this age of accountability is that the instructional practices being invested in will have a strong scientific evidence-base behind them. This focus on identifying evidence-based practices (EBP) has led to a flurry

of research in the education field, with rigorous studies being conducted to examine the efficacy of a wide range of educational practices, hence-forth referred to as programs in this proposal. This work is critical to the field, however simply knowing whether a program is efficacious is not enough to ensure that said practice will "translate" into positive outcomes in the real world of schools, teachers, and students (Coffey & Horner, 2012; Durlak, 2010; Odom, 2008). This next step involves understanding the contexts and conditions in which these programs work, the populations for whom these programs are effective, and the variables that impact the delivery of an efficacious program, leading to positive outcomes. This is the focus of implementation science, and the focus of this proposal (Harn et al., 2013; Odom, 2008).

The ways in which a program is implemented influences its impact on outcomes. Both effective and ineffective programs can be implemented well or poorly; program efficacy and implementation processes should not be conflated. A program found effective in a research study may lead to negative results when implemented poorly. In another situation, the same program may be delivered as intended but result in poor or negative student outcomes. Why might this occur? There may be a mismatch between the program's research-specified student population and the current school population, or the features identified from the research study may not be actual active ingredients of the program, or other contextual factors may need to be considered in selecting an EBP that will be effective in a specific school (i.e., mobility, ELL status). All of these issues, and more, have been reported in studies of implementation and impact obtained results (Durlak, 2010; Fixsen et al., 2005). Without question, implementation matters; the challenge is determining how to support effective implementation.

Improvements in educational outcomes are predicated, then, on two types of variables: intervention and implementation variables (Fixsen et al., 2005). The EBP movement has focused on identifying the efficacy of interventions, however the research examining how to effectively implement these programs is far more limited. Implementation is indelibly linked to complex contextual variables (e.g., school structures, resource allocation, teacher background and skills, student academic, behavioral, and emotional needs), which are harder to measure and define than intervention variables. As such their impact on student outcomes, while often recognized and discussed, is under-investigated (Fixsen et al., 2005; Webster-Stratton et al., 2011). Yet understanding what program works for whom, in what contexts, is critical for moving beyond basic program development to the sustainable implementation of effective EBPs that lead to improved outcomes.

There is a critical need for the field of education to attend to, define, and measure implementation variables (Coffey & Horner, 2012; Durlak, 2010; Flay et al., 2005; Harn et al., 2013; Odom, 2008; Odom, Hansen et al., 2010). This is the focus of this proposal. Toward this purpose, this chapter is broadly organized around several key ideas. After discussing what is generally known about implementation and synthesizing seminal articles on this topic, this literature review will make the case for a multidimensional and contextual approach to measuring implementation by summarizing what is known about: a) what elements of implementation are important and how they relate to each other; b) how to measure implementation (i.e., frequency, measurement approach); and c) the differential relationship of various aspects of implementation to student outcomes.

What Is Implementation?

Implementation, or the delivery of services in a given context, is a complex process. It is inherently ecological, impacted by interactions between multiple levels of systems, time factors, program characteristics, and stakeholder variables (Durlak, 2010). There is growing recognition of this complexity, however the lack of an agreed upon definition and/or measurement standards persists. Recently, there have been a number of well-cited literature reviews spanning a wide array of disciplines and literature bases (e.g., education, mental health, prevention science) that provide an overview of the way implementation is conceptualized and evaluated across these varying fields (e.g., Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003; Fixsen et al., 2005; Greenberg, Domitrovich, Graczyk, & Zins, 2005; Greenhalgh et al., 2005; Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993; Odom, Hansen et al., 2010; O'Donnell, 2008; Stith et al., 2006). Several of these reviews also advance important theoretical models of the factors that influence the implementation process (Durlak & DuPre, 2008; Fixsen et al., 2005; O'Donnell, 2008). Collectively, this literature base offers several theoretical frameworks for defining implementation, and makes important recommendations about the practical approaches needed to support the process and evaluation of implementation to improve student outcomes. In this section, an overview of three critical models of implementation will be provided, followed by a synthesis of recommendations and findings.

Fixsen et al. (2005). In their seminal review, Fixsen and colleagues examined the state of the science of implementation across multiple human services fields (i.e., mental health, juvenile justice, education, early childhood education, and social, employment,

and substance abuse services). This monograph has been incredibly influential not only for its thorough review, but also for its integration of broad system-level considerations (i.e., top-down variables, state-level policies, funding, theory, etc.) and microsystem level considerations (i.e., bottom up variables, program variables, provider training and delivery, etc.) of the variables that impact intervention implementation (Durlak, 2010; Odom, Cox, & Cook, 2013).

Overview of findings. Fixsen et al. examined 1054 articles that focused on implementation. Of those, 743 met study design criteria (i.e., literature review, empirical analysis, or meta-analysis of implementation variables), and 377 were deemed "significant"; only 22 were meta-analyses or studies that experimentally manipulated implementation variables (p. 68-69). This review highlights the dearth of quality research across multiple fields that examine the complex variables involved with implementation, and the lack of well-defined terms, measures, or procedures for studying these variables. Despite these limitations, Fixsen et al. note that, given the diversity of programs, variables, and fields examined in this review, the significant degree of convergence actually provide strong support for their model of implementation (2005).

Approach to implementation. In summarizing this literature base, Fixsen and colleagues highlight the fact that implementation of *any* program across *any* human service field is a complex process that is influenced by variables at multiple levels of the organization. Fixsen et al. operate from a purposefully flexible view of implementation to target the full range of programs and content being studied in the implementation literature, defining it as "a specified set of activities designed to put into practice an activity or program of known dimensions" (2005, p. 5). They situate this work in the

context of the EBP movement, emphasizing the importance of implementation variables in addressing the "science to service gap". Fixsen and colleagues note the importance of differentiating between intervention and implementation variables, both of which include different processes and outcomes, when examining any program. Intervention variables include program characteristics, the active core ingredients, and effectiveness outcomes. In contrast, implementation variables address issues such as adherence to the program, quality of delivery, and contextual factors that impact the delivery process. Implementation outcomes include changes to practitioner behaviors, changes to organizational structures, and changes to relationships with participants or consumers. Positive outcomes should only be truly *expected* in cases where both intervention and implementation variables are accurately applied, or where, effective practices are fully implemented. One of the key ideas discussed in this monograph was that implementation variables should not be assumed any more than intervention variables, and as such should be measured, analyzed, and reported in research findings (2005; p. 4).

Implementation descriptions. Fixsen et al. (2005) also provide an in depth description of the factors demonstrated to effect implementation in their review. Key features, such as active involvement and training by program developers, and the importance of long-term multilevel implementation supports (e.g., skill-based training, coaching, staff and program evaluation) are highlighted as effective implementation strategies. While much of the evidence supporting these recommendations was found in research examples showing what *does not* work (e.g., passive provision of program information), the consistency of these findings was notable across time, domains, and programs. Despite this consistency, these patterns also underscore the need for high

quality research examining implementation variables, and their relationship with outcomes. They noted that this gap in the research is particularly concerning in terms of examining the broader organizational and system-level variables that impact implementation, and in looking at implementation variables over time.

Fixsen et al. (2005) also note "the obvious," that implementation takes place in community contexts, and detail several important community variables associated with implementation. They highlight the widely agreed-upon importance of addressing, and measuring, community readiness and staff buy-in when beginning the process of implementing a new program. Several factors are identified as important to supporting this process (e.g., administrative supports, positive staff attitudes and beliefs), however here too, their literature review returned scant findings of studies that empirically examined these variables in initial implementation or evaluation.

Implementation frameworks. In addition to the discussion on the characteristics involved in successful implementation, Fixsen and colleagues (2005) also developed several important implementation models and frameworks based on their review. Of particular importance for the proposed study are their framework of core implementation components, and their overview of the stages of implementation.

Core components. Fixsen et al. (2005) identify six core components as necessary for the effective implementation of EBPs: (a) staff selection, (b) preservice and inservice training, (c) ongoing consultation and coaching, (d) staff and program evaluation, (e) facilitative administrative support, and (f) systems interventions. Staff selection involves deciding who will deliver a chosen practice, what their qualifications will be, and how they will be selected. Preservice and inservice training involve initial professional

development aimed at increasing staff awareness, knowledge, understanding of program theories and philosophies, and providing initial practice opportunities, however it should be noted that these types of trainings are considered to be ineffective implementation strategies when used independently. Coaching and consultation serves as the primary mechanism through which practitioner behaviors are shaped and changed to support appropriate delivery of the program. This component integrates "on the job" training, practice opportunities, and support and should be offered across each stage of implementation. Staff and program evaluation describes the process of assessing the use, delivery, and outcomes of the specified program at various points in time. Administrative support involves leadership practices that facilitate the process of implementation, such as data-based decision making, staff organization, and the provision of implementation goal-oriented supports. Closely related to administrative supports, systems interventions involve the system-level strategies that provide the financial, organizational, and human resources necessary to support practitioners in program implementation.

These six components are highly related, and the authors (2005) speculatively recommend viewing them as compensatory, with weaknesses in one area being addressed by strengths in another. This compensatory process, while a strength when considering the complex nature of these integrated systems, also presents a challenge in that changes in one component or process requires adjustments to the other components as well to ensure that effective implementation is maintained over time. This requires consistent measurement and evaluation of these implementation processes (Fixsen, Blasé, Naoom, & Wallace, 2009; Fixsen et al., 2005).

Stages. As was noted in the previous chapter, Fixsen and colleagues found evidence of six distinct phases in the implementation process: (a) exploration, (b) installation, (c) initial implementation, (d) full operation, (e) innovation, and (f) sustainability. During the exploration phase, key stakeholders (e.g., administrators, district-level policy makers) identify the need for a program in their setting, and begin the process of collaborating with program purveyors (e.g., researchers) to assess the needs of the school, the fit between the school and the program, and to prepare the school for installation. Program installation involves preparing the school for initial implementation, ensuring that start up resources (e.g., staff, materials, training provisions) are properly allocated, so that necessary supports are available once the program is put into place. Initial implementation is the start of program delivery, while full operation is considered underway once program delivery is fully integrated into typical practice across the school. The innovation phase occurs when the staff moves beyond basic program delivery, and begin to adapt the program to meet and support contextual needs specific to that school setting. Finally, the sustainability phase involves maintaining systems (e.g., staff training, hiring practices, funding) that ensure that the program continues to be implemented effectively over time (2005, p. 15-17). As with the core components, Fixsen et al. suggest that these stages of implementation are interrelated, with evidence that early implementation processes impact later implementation outcomes, though specific variables, such as administrative supports, that were found to be differentially important during early phases were not necessarily significant or related to outcomes in later phases. Fixsen and colleagues stress the fact that these stages, while distinct, should be thought of as recursive, as different contextual variables (e.g., high rates of staff turn-

over) may result in programs moving from sustainability or full operation to initial implementation for a period of time while staff receiving training and support (Fixsen et al., 2009; Fixsen et al., 2005).

Both of these particular frameworks (core components and implementation stages) hold important implications for issues related to measuring and evaluating implementation variables. Understanding and evaluating where an organization is in the process of implementation (i.e., stages), and which elements of the organization are driving that process (i.e., core components), is critical to the process of measuring implementation. There is limited research on these systems and processes as a whole, given the dynamic process of implementation, however thinking about these components and stages as aggregate units may offer important insights when measuring implementation.

Implications. Fixsen and colleagues (2005) have presented the field with a comprehensive review of the current state of implementation research, a description of the variables that impact implementation processes and outcomes, an outline of effective implementation practices, and the role of multiple organizational levels (i.e., practitioner and broader system-level pieces) in the implementation process. The synthesis and frameworks they created based on this evaluation have major implications for the field, particularly in terms of developing common language and models as starting points for the study of this complex topic. As such, the elements of Fixsen et al. serve as the starting point for the proposed study. There is a need, however, to hone in on specific issues related to education, and to the measurement of implementation in specific

educational contexts. For that, an analysis of two additional reviews of the implementation literature base is necessary.

O'Donnell (2008). In her 2008 review of K-12 core curriculum intervention research, O'Donnell analyzed quantitative studies that specifically examined the relationship between implementation and outcomes. She focused on primary research studies that examined the efficacy and effectiveness of K-12 curriculum interventions across core subject areas (i.e., reading, math, science, social studies) that could be implemented in individual classrooms. With this approach, O'Donnell specifically excluded studies that focused on the implementation of whole-school programs (e.g., SWPBIS, RTI), which incorporate many of the broader system-level variables Fixsen et al. (2005) described, and as such narrowed her focus to the relationship between implementation variables and outcomes in specific classroom practices.

Overview of findings. O'Donnell (2008) reviewed 120 publications examining the implementation of K-12 core intervention programs where only 23 were empirical studies that met review criteria. Five out of those 23 studies met full inclusion criteria by quantitatively examining the relationship between implementation of core curriculum interventions and outcomes (2008; p. 37). Given this data set, the patterns drawn from these findings are a bit limited, however O'Donnell draws from the larger examined literature base and provides an insightful overview of several important issues related to the measurement of implementation variables, and their relationship to student outcomes in educational contexts.

Approach to implementation. O'Donnell (2008) approaches implementation with a traditional focus on *fidelity of implementation.* While the definition of this term is not

widely agreed upon (e.g., Odom, Hansen et al., 2010), she defines it as "the determination of how well an intervention is implemented in comparison with the original program design during an efficacy and/or effectiveness study" (2008; p. 33). Her definition draws from a review of both the health and education fields, through which she concludes that the majority of studies describe implementation as a synonym for adherence to program components. O'Donnell notes the limited nature of this perspective, and encourages a more multidimensional approach to measuring implementation, which will be described in more depth later in the chapter. O'Donnell situates this work in the context of increasing interest in the development of EBPs in education, and outlines the need for increased measurement of implementation variables beyond the initial research phases (efficacy), and appropriate tools with which to do this work.

O'Donnell (2008) focuses on implementation measurement in two specific research contexts (i.e., efficacy and effectiveness studies), however overlooks other implementation stages (e.g., full operation, sustainability; that should be examined in implementation research (Fixsen et al., 2005). While this is a limitation, this approach of examining and measuring implementation differently for different *evaluation purposes* has major implications for the field. As O'Donnell indicates, more research is needed to examine the specific types of measures that should be used in specific contexts.

Measurement issues. O'Donnell (2008) describes several measurement considerations, though given the limited sample size she broadened the review to include studies that specifically provided recommendations on how to measure implementation. First, O'Donnell describes the wide range of measures used to collect implementation

data, including surveys, self-report, interviews, and direct observation methods such as video, audiotape, and live observation. She notes that these measures were used to capture a range of implementation variables, such as adherence to program procedures, quality of delivery, and teacher and student responsiveness. O'Donnell also highlights the importance of evaluating these measures in relation to student outcomes. In this review, only 5 studies examined this question, however each found higher outcomes when implementation was higher; the majority of these studies used limited models to measure this relationship.

Finally, and most importantly for the proposed study, O'Donnell (2008) examines the theoretical frameworks that were used to guide the measure of implementation. She notes that there was a range of approaches, with two studies using no theoretical frameworks, and the other three using various approaches that highlighted the measurement of contextual variables, such as teacher adaptation, and buy-in. While it's difficult to generalize from such a small sample, this idea of assessing the use of theory to drive implementation measurement is important. O'Donnell notes that program theory, or the theoretical framework used to design the intervention, should directly impact the implementation variables that are measured, and that this process of identifying specific implementation criteria should be developed a priori to intervention delivery. This closely aligns with the notion of identifying the active core ingredients of EBPs discussed in Chapter One. While this is (or should be) a critical step in the study and measurement of EPBs, this may represent a conflation of the intervention and implementation variables discussed by Fixsen et al. (2005). Ideally, this work would be done during program development/efficacy research, and have a limited, or perhaps follow-up role (e.g., How

does adherence to core ingredients maintain or shift in real world applications, and how does that relate to outcomes?) in effectiveness research and other studies focused on later implementation stages. This is rarely done in practice (Durlak, 2010; Harn & Parisi, 2013), and as such this recommendation highlights the difficulty involved in measuring all of the complex tasks involved in implementation. Despite these complexities, the notion of using theory to drive the measurement of implementation-specific variables (e.g., the importance of structural vs. process dimensions, level of adherence required, importance of quality vs. dosage) is still an important recommendation, and one that will be returned to later in this chapter.

Implications. O'Donnell's (2008) literature review has several limitations (e.g., sample size, conflating intervention and implementation variables), however the following discussions are key to supporting the implementation of EBPs beyond the research study: (a) the importance of program theory when developing implementation measures, (b) the application of different measures and tools to different implementation contexts, and (c) the importance of developing multifaceted implementation measures. O'Donnell also applies this discussion of implementation to classrooms specifically, and notes the importance of considering context variables, such as the relationship between teachers and curriculum. This review, however, while highlighting important issues for considerations when measuring implementation variables, falls short in terms of offering an applicable conceptual model of that process. For that, it's necessary to turn to a third important review of the literature.

Durlak & DuPre (2008). Durlak and DuPre (2008) analyzed the relationship between implementation and outcomes across over 500 quantitative studies from a wide

range of human service fields (i.e., mentoring, after-school programs, drug prevention, health promotion and prevention, mental health). They also conducted a secondary analysis of both quantitative and qualitative studies to examine the contextual factors that affect implementation. Like Fixsen and colleagues (2005), Durlak and DuPre purposefully cast a wide net to gain insight into the complex factors impacting implementation and its measurement across the human service field.

Overview of findings. In their primary analysis, Durlak and DuPre (2008) examined the relationship between intervention implementation and participant outcomes across 542 interventions. The majority of these (n = 483) were studies that were summarized in five meta-analyses, however they also examined 59 additional quantitative studies that specifically measured this relationship. Data from the meta-analyses provided strong support of the fact that implementation influences outcomes, with the majority of studies indicating that programs with higher levels of implementation led to increased participant outcomes. Implementation was found to be one of the most important predictors of program outcomes, with stronger implementation scores leading to mean effect sizes that were two to three times higher than programs with lower implementation.

In their examination of the 59 primary studies, Durlak and DuPre (2008) noted similar results, with 76% of studies reporting a significant positive relationship between implementation, and at least half of the program outcome measures. In the remaining studies, there often was not enough variability in measured implementation to detect *any* relationship (i.e., implementation was consistently high/low). Durlak and DuPre note the consistency of these findings across a wide range of studies, interventions, and domains,

indicating that implementation matters for improving participant outcomes. In their secondary analysis, Durlak and DuPre (2008) examined an integrated quantitative and qualitative literature base to determine the contextual factors found to impact implementation processes. They analyzed 81 studies, and identified 23 specific factors. These factors were grouped across five ecologically oriented implementation categories (i.e., community factors, provider characteristics, innovation characteristics, prevention delivery system, prevention support system). Broadly speaking, these factors confirm the fact that implementation is a complex process, impacted by multiple levels (e.g., individual, organization, community), and that there is significant interaction and overlap amongst many of these factors, making them difficult to isolate and measure. Factors were only included if they were related to implementation in at least five studies, and if the findings across those studies were consistent with the most rigorous study in that group. More rigorous factor analysis (e.g., quantitative confirmatory factor analysis) may be necessary to confirm these findings, however Durlak and DuPre compared these findings to several other implementation reviews, including Fixsen et al. (2005), and note that there is a great deal of convergent validity across these contextual factors.

Approach to implementation. Durlak and DuPre (2008) define implementation as "what a program consists of when it is delivered in a particular setting" (2008; p. 329). They situate this review in a capacity perspective, where capacity refers to the entire process of diffusing effective programs from research into long-term sustainable practice. They note that there are several phases to this process, *after* a program has been developed and identified as effective (e.g., evidence-based): (a) dissemination, whereby a program is introduced to potential communities or users; (b) adoption, or the time when

said community decides to attempt to use the program; (c) implementation, or how well the program is delivered over a set "trial period"; and (d) sustainability, or the program's long term maintenance. While these phases have some overlap with the Fixsen et al. (2005) stages of implementation, this isolation of implementation as an individual component of the broader process of diffusion is unique. Durlak and DuPre also use this view of implementation as a unique stage within the diffusion process to isolate specific measureable variables that can be independently evaluated. These variables hold important measurement implications for the field, and as such are discussed in detail below.

Aspects of implementation. Durlak and DuPre (2008) also emphasize the challenges related to measuring implementation that arise out of the lack of clear terms or consensus around definitions. As such, they provide an in depth overview of the multicomponent aspects of implementation that are referred to throughout the literature, drawing on the work of Dane and Schneider (1998), and integrating additional components based on their own reviews of the literature. Durlak and DuPre note that some of these components can have slightly different meanings depending on context, but they contend that these eight aspects of implementation (i.e., adherence, dosage, quality, participant responsiveness, program differentiation, monitoring control conditions, program reach, adaptation) capture the broad range of variables that researchers generally focus on examining when they measure implementation. (1) Adherence, also known as fidelity to treatment or integrity, refers to the extent to which a program is implemented as intended. (2) Dosage, or quantity, addresses the amount of the prescribed program that is delivered. (3) Quality refers not to program content, but to

how well it is implemented, which can encompass measures of implementation clarity or correctness, perceived effectiveness, or more qualitative variables such teacher enthusiasm, or emotional connection to students. (4) Participant responsiveness measures the degree to which participants respond to the implementation of the program (e.g., teacher enthusiasm for an intervention, student participation or progress). (5) Program differentiation refers to the extent to which the critical components of the program (e.g., theoretical framework, specific practices) can be measured as unique and distinguishable from comparison programs (Dane & Schneider, 1998; Durlak & DuPre, 2008; O'Donnell, 2008). (6) Monitoring control or comparison conditions refers to the comparison of treatment and control groups in an effort to measure program effects. (7) Program reach describes the rates of participation across populations of participants, and the scope of the program during delivery. (8) Adaptation is a highly contentious aspect of implementation, and refers to changes made to specified programs during actual delivery (Durlak & DuPre, 2008; Odom, Hansen et al., 2010). In clarifying this view of implementation as a broad, multifaceted construct, Durlak and DuPre provide a critically important model for researchers interested in measuring these related, but distinguishable components of implementation.

Measurement applications. Durlak and DuPre evaluated which aspects of implementation were measured across this literature base, how those aspects were assessed, and each aspect's relationship to outcomes. Overwhelmingly, these studies measured individual aspects of implementation (n = 41), though 15 studies did examine two aspects, and three studies examined three (e.g., fidelity, dosage, adaptation). With the exception of two studies that examined only program reach, all studies examined

structural dimensions of implementation (i.e., adherence, dosage), and very few studies included other aspects, with six including a measure of quality, four examining reach, and three including adaptation. These aspects of implementation were measured primarily through self-reports or direct observation, and while Durlak and DuPre note that direct observation is more likely to be linked to outcomes, they also highlight the fact that any check of aspects of implementation during program delivery can be useful in identifying areas of concern. All but one study had separate measures for each aspect of implementation that was evaluated, though several studies had multiple measures of specific aspects.

Finally, the majority of studies, including the three that measured adaptation, found significant positive relationships with the measured aspects of implementation and outcomes. While this is a clear indication of the fact that implementation impacts outcomes, this review also identified other important elements of the implementation process. Durlak and DuPre (2008) found significant variability in implementation levels across studies, with common ranges of 20 – 40 percent. They also found a wide range of implementation levels that were associated with positive outcomes, with most studies reporting positive results when implementation levels were between 60-80%, and very few studies reaching levels higher than 80 percent. Durlak and DuPre, while emphasizing the fact that these studies provide clear empirical evidence that implementation level affects outcomes, also discuss this implementation variability as an important contribution to the "fidelity versus adaptation" debate. They note that this finding provides evidence of the fact adherence and adaption can co-occur in real world settings, and that both may be necessary to measure and support.

Implications. Durlak and DuPre's (2008) model of implementation as an overarching construct with related but distinct subcomponents, and the measurement approach described above provides a powerful tool for implementation researchers. This approach bridges many complex factors across the broad implementation literature base, and has implications for the development of a clear measurement process that can be applied across multiple studies and implementation frameworks. Their review provides important support for the conclusion that implementation does matter, however, as they note in their discussion, there are still many unanswered questions and considerations that need to be empirically examined. Many of the questions outlined in their conclusion are incorporated into the proposed study, such as the recommendation that researchers examine: (a) multiple aspects of implementation in relation to program outcomes; (b) the role of time; and (c) stage of implementation. These will be described in detail later in this chapter, with additional information from other sources, however it is important to acknowledge their root in this seminal review.

Assessing Implementation: Measurement in a Context

The implementation literature base highlights three underlying features to consider when measuring implementation: (a) it impacts outcomes, (b) it is a complex process impacted by contextual variables, and (c) it requires a nuanced measurement and evaluation approach to truly understand its role. In other words, implementation is not simply program delivery, and cannot be measured as such. It involves what is delivered, as well as how it delivered, by whom, to whom, as well as when, where, and why a program is delivered. It also involves the broader systems that support that delivery process, and actual student and practitioner response to program delivery. Each of these

components may need to be measured to truly understand implementation, and how it links to outcomes. This points to the need to conceptualize the process of assessing implementation as measurement in a context.

This concept of measurement in a context entails a comprehensive and systematic approach to identifying the variables involved in implementation, and determining which variables should be measured at a given implementation point. As O'Donnell (2008) highlighted, the critical variables during efficacy research (e.g., program theory, adherence to program structure) may be very different from those involved during dissemination or scale-up (e.g., adaptation, technical assistance or training). Durlak and DuPre (2008) and Fixsen and colleagues (2005) also note that implementation involves multiple components and factors at any point in time throughout the process. Measuring implementation must therefore involve measuring multiple aspects of the implementation process at different times. This requires a view of implementation as a multifaceted construct.

Considering implementation as a multifaceted construct. As Dane and Schneider (2005), Durlak and DuPre (2008), O'Donnell (2008) and others have highlighted, implementation is a broad and multifaceted construct with as many as eight subcomponents or aspects. While some of these aspects of implementation have a robust body of research behind them (i.e., dosage, adherence), others have received limited attention (i.e., participant responsiveness, program differentiation) and there is growing recognition that these aspects of implementation need to be studied concurrently, to determine their interactions as part of the broader construct of implementation, and their relationship with intervention outcomes (Durlak & DuPre, 2008).

Drawing from the public health field, Mowbray, Holter, Teague, and Bybee (2003) recommended that implementation variables be considered across two broad dimensions: structure and process. Structural variables relate to the quantity of the specific program, while process variables relate to the quality of the implementation (Mowbray et al., 2003; O'Donnell, 2008; Power et al., 2005). Structural aspects of implementation involve the specific framework for program delivery, and as such includes measures that target how *many* of the structural components of a program are implemented (e.g., how much of the program or content was implemented). O'Donnell (2008) categorized dosage and adherence as structural aspects of implementation. Process-oriented aspects of implementation, on the other hand, involve the way services are delivered, and therefore include measures of how *well* the instructional program is implemented (e.g., how was the delivery process). O'Donnell categorized quality of delivery and program differentiation as process aspects of implementation. Participant responsiveness contains elements of both process and structural aspects of implementation.

Durlak and DuPre's (2008) additional subcomponents of implementation have not been explored in this vein, however speculatively, these aspects fit these same dimensions. Program reach is related to measuring how various programs target specific populations, and therefore assesses implementation process. Monitoring comparisons involves assessing the ways services are delivered in contrast to typical or control programs, and as such is a structural element. Adaptation, like participant responsiveness, seems to contain both structural and process oriented elements as teachers may make changes to both the structural (e.g., amount of instruction, modifications to activities or

procedures) and process (e.g., student response formats, language or enthusiasm) components of a program.

Measurement applications. In education we have typically focused on either the structural or process dimensions, and often examine only adherence. In a recent review of the school psychology literature base, Sanetti, Gritter, and Dobey (2011) found that only about half of studies reviewed over thirteen years included an implementation measure, and that almost all focused solely on structural aspects. Durlak and DuPre's (2008) findings mirror this, where the majority of studies included in their review took a unidimensional approach to implementation, and almost all focused on structural aspects of implementation. O'Donnell (2008) found that included studies used a unidimensional measure of implementation, however she found a split between studies that measured process and structural dimensions. Despite these limited approaches, several other researchers have echoed O'Donnell's (2008) recommendation to examine implementation as a multifaceted construct (e.g., Harn et al., 2013, Odom et al., 2010; Power et al., 2005; Webster-Stratton et al., 2011). This next section will review studies that have examined the relation of different aspects of implementation to each other. Later, a review of the few studies that have also examined how various aspects of implementation relate to student outcomes will be provided.

Examining the relationships between various aspects of implementation. Knoche, Sheriden, Edwards, and Osborn (2010) examined the effects of teacher training on multiple aspects of implementation in the delivery of an early childhood intervention that targeted school readiness. Knoche et al. examined the relationships between adherence, dosage, quality, and both structural and process-oriented aspects of participant

responsiveness, while also comparing treatment and control groups in an effort to examine program differentiation. Results indicated that quality and adherence were generally highly correlated, though these authors stress that each measure contributed unique information to the overall picture of implementation obtained in the study. Quality was also found to be significantly related to dosage, and to the process-oriented measure of responsiveness, though only with the teachers considered as having more training. Adherence, on the other hand, was found to be significantly associated with structural elements of responsiveness. Knoche et al. offer this approach up as a model for the examination of structural and process elements of implementation, noting the comprehensive and complex insight they gained into the intervention and implementation process by taking this multifaceted approach.

Domitrovich, Gest, Jones, Gill and Sanford DeRousie (2010) also examined multiple aspects of implementation in their study of a comprehensive preschool curriculum in Head Start classrooms. Domitrovich et al. included measures of dosage and adherence, as well as measures of child engagement and teachers' abilities to generalize curricular content outside of the scripted intervention, which they considered to be important components of quality of delivery. While these authors did not address the relationships between these aspects directly in their study, they note that their measures of adherence and child engagement were strongly correlated. These authors note that this should not be interpreted as a conflation of one construct, highlighting the difference between teachers' adherence to intervention protocol and student's engagement with the materials. They speculate that this relationship may be due more to the nature of the intervention's active core ingredients, which specifically aimed to maximize engagement.

In their study of the yearlong implementation of a comprehensive integrated preschool curriculum, Odom et al. (2010) examined the relationships of a structural (dosage), process (quality), and multicomponent (composite) measure of implementation. They found strong, significant correlations between the mean ratings of each aspect of implementation within and across curricular areas (e.g., math, literacy, social skills). These associations were always higher between a singular dimension and the multicomponent measure (e.g., literacy quality and literacy multi-composite), as the multicomponent measure was a composite, but the strength of these associations was high and strong across the board.

Interestingly, Hamre et al. (2010) found no association between multiple measures of dosage, adherence, and quality aspects of implementation in their study of a supplemental literacy and language development curriculum. This was true even when examining multiple measures of the same aspect. Hamre et al. examined two dosage measures, one that relied on teacher self-report, and one that relied on direct observation, and found very weak associations between the two. This same pattern was found between two distinct measures of quality, one that targeted teacher's language modeling, and one that targeted literacy focus. These authors suggest that these findings indicate that these measures clearly capture unique aspects of the broader implementation process.

Collectively, this small but important research base demonstrates the range and variability in how different types of implementation measures relate to one another, and makes a strong case for the need to examine implementation as a multifaceted construct (Durlak, 2010). Despite the sometimes contradictory findings presented in these articles, all stress the importance of including multifaceted measures of implementation as well as

the need for further research (Durlak, 2010; Durlak and DuPre, 2008; Harn et al., 2013, Odom et al., 2010; O'Donnell, 2008).

Measurement Considerations

Given the important role of measurement in understanding implementation and its relationship to program outcomes, the proposed study aims to address this gap by focusing on the following measurement considerations: (a) what to measure (e.g., which aspects or dimensions of implementation are important for a given implementation evaluation); (b) when to measure it (e.g., stage of implementation, stage of research, as well as frequency); and (c) how to measure it (e.g., observation approach, types of measures). Each of these measurement considerations will be discussed next. Specifically, this section will first examine delivery variables (i.e., what to measure). This will be followed by a review of a technical observation measurement guide (Chomat-Mooney, 2008) with important implications for the remaining two measurement considerations, including relating implementation to outcomes.

What to measure. In their initial discussion of a multidimensional view of implementation, Dane and Schneider (1998) recommend that researchers include all five of their aspects of implementation, when appropriate. Durlak and DuPre (2008), in extending the number of aspects to eight, are less prescriptive, but do recommend that researchers analyze multiple aspects depending on the context of a given study. The examples of a multifaceted approach cited in the previous section highlight examples of researchers engaging in this process, as all of the included studies measured more than one aspect of implementation (Domitrovich et al., 2010; Knoche et al., 2010). While this is certainly an improvement over the typical unidimensional measurement approach (e.g.,

Durlak & DuPre, 2008; O'Donnell, 2008; Sanetti et al., 2011), these examples also highlight the fact that measuring all eight aspects of implementation may not be practical, or even possible when one considers issues of efficiency, and the resources and measures available. This then raises the question of which aspects to include, and how to make that decision.

Deciding which aspects of implementation to measure is the starting point in this process, and these decisions should be made a priori, based on thoughtful considerations of the program and research context. Researchers should start with a firm understanding of the underlying mechanisms of the program, and the specific delivery context, and then use that to drive the development of their implementation measurement framework (e.g., which aspects of implementation to include, and why, how to measure those aspects) (O'Donnell, 2008; Mowbray et al., 2003). This involves examining specific delivery variables (i.e., program content, program context) from a measurement perspective, as these variables have important implications for what should be measured. Specifically, program content should be examined to determine which content variables are actually leading to intended changes in student outcomes. From a measurement perspective, this involves considering the program's (a) active ingredients, (b) adaptability, and (c) program theory. Program context should be examined to determine the role of variables in the context of delivery that contribute to those same changes. There are many important contextual variables that impact implementation (e.g., community context, organizational capacity, structural support systems), however given the proposed study's focus on classroom implementation, only contextual factors related to teacher variables and program-context alignment will be examined in depth for the purposes of this study.

Each of these content and contextual variables will be discussed from a measurement perspective next.

Content considerations. Typical approaches to measuring implementation often focus on monitoring the delivery of program content to ensure that a program is delivered as intended by program developers (i.e., adherence). This is most often done for "curriculum-dependent" programs with a packaged scope and sequence of procedures and activities (e.g., *Reading Mastery*), though program content can also be monitored for "curriculum independent" programs that are independent of a set curricular area, topic, or packaged program (e.g., explicit instruction). While these basic content measures provide information about teachers' adherence to specific elements of programs, this limited approach to measuring content does not always capture the more nuanced content variables that impact a program's implementation.

Active ingredients. It is often assumed that programs and practices that are considered evidence-based have been empirically examined to determine the components that function as the change mechanisms in the intervention. This assumption applies to both "curriculum-dependent" and "curriculum-independent" programs, though the common inclusion of adherence measures in curriculum-dependent programs may provide a misplaced sense that active ingredients are known. In the context of measuring implementation, both these types of programs should be evaluated comprehensively to determine the actual change mechanisms that support the intervention's ability to positively impact student outcomes. This requires the understanding, and *measurement*, of a program's active ingredients (Durlak, 2010; Harn & Parisi, 2013; Harn et al., 2013).

This work involves systematically analyzing which components of an intervention are critical to achieve the determined effect, as well as the dosage and duration required, and delivery features that support effective implementation (e.g., Do all activities and procedures need to be completed? For the entire prescribed time? What teacher variables and behaviors are important for delivery? Do all students need to receive the same program components, or do different groups have different needs?) (Durlak, 2010; Harn & Parisi, 2013; Harn et al., 2013). The majority of interventions are developed as packaged programs based on researchers' integration of theory, research, and practice into a range of activities and procedures (Harn et al., 2013). Efficacy studies that examine these programs typically focus on the intervention as an integrated package, and their evidence-base is predicated on the delivery of the full package under ideal research conditions. There is little evidence that the type of component analysis required to effectively identify active ingredients has been conducted for the majority of programs touted as evidence-based (Durlak, 2010; Harn & Parisi, 2013; Harn et al., 2013).

In addition to considering the active ingredients of a program, researchers should also consider program theory, or the intervention's underlying theoretical framework, when measuring implementation (Mowbray et al., 2003; p. 315). Active ingredients should be based on this theory, however the underlying theory of action may involve much more complex, nebulous processes (Harn et al., 2013; Mowbray et al., 2003). Implementation measures should therefore be based on a clear understanding of program theory, as adherence to the theoretical underpinnings of an intervention is what should be assumed to drive student outcomes, rather than adherence to basic procedures or activities. These may be highly related, however this is not always the case, particularly

with more complex interventions such as curriculum-independent programs (e.g., explicit instruction, SWPBIS), or programs that involve practitioner decision-making, adaptations based on contextual variations, or interventions that must be individualized for participants with a wide range of varying needs and abilities (Mowbray et al., 2003; p. 326).

Fixsen and colleagues (2005) highlighted the important distinction between intervention and implementation variables. The work of identifying active ingredients and program theory focuses on intervention variables. This work should ideally be conducted during efficacy trials, as these variables should be clearly identified to guide implementation research and identify which aspects of implementation are important for a given program (e.g., Which variables must be adhered to? Where is there room for adaptation?) (Fixsen et al., 2005; Durlak, 2010). As this is rarely the case, however, many researchers are beginning to acknowledge implementation research as a platform for evaluating these active ingredients systematically, as a key factor in determining the effectiveness of implementing various EBPs in "real world" settings (Fixsen et al., 2005; Flay et al., 2005; Harn et al., 2013). As such, if the active ingredients and/or program theory for a given program have not been systematically identified, comprehensive component analyses that include a full or wide range of implementation measures (e.g., all aspects of implementation) may be warranted.

Adaptation. Durlak (2010) also notes the important role of active ingredients in understanding program adaptation. Adherence and adaptation are often viewed as polar opposites, particularly by researchers focused on strict adherence to program components for the purpose of documenting internal validity. Yet in their review of teachers'

adherence to curriculum guides for substance abuse prevention programs, Ringwalt et al. (2003) found widespread use of adaptation, and noted that "some measure of adaptation is inevitable and that for curriculum developers to oppose it categorically, even for the best of conceptual or empirical reasons, would appear to be futile." (2003; p. 387). Fixsen and colleagues (2005) challenge the notion that adaptation is inevitable, noting that when strong adherence to program's active ingredients is the goal, flexibility around other procedures is acceptable, but should not be considered as adaptation as the program's core components remain the same. Durlak and DuPre (2008), on the other hand, found that both adherence and adaptation aspects of implementation could be positively related to participant outcomes, depending on context. These authors note that some clearly scripted and highly prescribed programs may inherently support stronger adherence, while other less standardized or curriculum-independent programs may not.

Webster-Stratton et al. (2011) provide an interesting example of how adherence and adaptation aspects of implementation can be integrated into program development. They note that their evidence-based classroom management intervention the Incredible Years (IY) is flexible and principle-driven, rather than a fixed-dosage or curriculum-driven program (Webster-Stratton et al., 2011; p. 512). This allows coaches and professional developers to provide contextualized, responsive training that supports teachers in implementing the IY program to match their context. Webster-Stratton and colleagues identify core components (i.e., active ingredients) that are required during all trainings, however they also note that program components such as pacing, dosage, level of support, examples, number of activities, and practice opportunities may be adjusted for different teachers and contexts. Supplemental materials and additional trainings are also

available for teachers working with culturally and linguistically diverse children and families, students with particularly challenging behavior, and for parent supports. This has huge implications for the long-term sustainability of this EBP, though there is limited discussion of how or when to measure these built in adaptations, or how they should be related to overall decisions around the implementation process.

While the "fidelity versus adaptation" debate often pits researchers squarely in the fidelity camp, this more nuanced view of active ingredients emphasizes the components of a program that must be adhered to while leaving room for practitioners to adapt programs to make them more contextually and culturally relevant to their settings and student populations. When active ingredients are clearly identified and tied to implementation measures, researchers, teacher trainers, professional developers, and coaches can support teachers in this process, rather than just insisting that de-contextualized intervention packages be implemented with strict adherence (Durlak, 2010; Durlak & DuPre, 2008; Harn et al., 2013). Both the adherence and adaptation aspects of implementation must be accurately measured, however, to ensure that their influence on the implementation process is accurately understood.

Context considerations. Multiple reviews of implementation research highlight the importance of contextual variables on the implementation process, and situate analysis of these variables in an ecological framework (i.e., Durlak & DuPre, 2008; Fixsen et al., 2005; Greenhalgh et al., 2005; Stith et al., 2006). A full review of these variables is beyond the scope of this paper (see Durlak & DuPre, 2008 and Fixsen et al., 2005 for comprehensive discussions of ecological factors), however several of these contextual factors (e.g., organizational capacity) have been regularly found to relate to

outcomes, and should be considered in implementation research. Given this study's focus on classroom-level instructional implementation, however, only those contextual factors related to teacher quality and program fit will be discussed from a measurement perspective.

Teacher quality. Teachers, as the leaders of classrooms, have a huge impact on classroom contexts. Several lines of research have examined basic teacher characteristics, such as teacher education, experience, and credentials. While some studies have found limited associations between teacher education and classroom quality, the literature base on these teacher characteristics should be considered equivocal at best, with recent studies indicating few if any systematic associations between teacher education, classroom quality, and student outcomes (La Paro et al., 2009; Lieber et al., 2009). Other more nuanced teacher characteristics, such as teacher attitudes and beliefs, also impact classroom contexts. These types of characteristics, while more difficult to measure, have also been linked to the quality of teachers' implementation of specific practices (e.g., Bambara, Nonnemacher, & Kern, 2009; Klingner, Ahwee, Pilonieta & Menendez, 2003; Lieber et al., 2009). While these teacher characteristics are often not examined directly in terms of associations with student outcomes, there is widespread agreement in the field that teacher-level variables are key factors to consider when looking to improve student-level outcomes, as teachers "mediate all relationships within instruction," (Cohen & Ball, 1999, p. 10; Fishman, Marx, Best, & Tal, 2003; Garet, Porter, Desimone, Birman, & Yoon, 2001).

Teachers may have an impact on the implementation process that extends or inhibits the actual program's reach. From a measurement perspective then, it is important

to understand the extent to which teachers influence implementation. Durlak (2010) emphasizes this point, problematizing the assumption that program implementation occurs outside of the context of a teacher's ability to provide effective instruction (p. 354). This underscores the importance of general teacher competency, and teacher effectiveness variables (e.g., behavior management skills, organizational skills, clarity of presentation, modeling, feedback processes, relationships with participants) that exist externally to a specific program. Decades of research supports the importance of these variables (e.g., Brophy & Good, 1986; Cohen & Ball, 1999; Fishman et al., 2003; Gage & Needels, 1989; Rosenshine, 1997; Simonsen, Fairbanks, Briesch, Myers, & Sugai, 2008), and Durlak notes that implementation research should also account for the way overall teacher quality impacts the implementation of specific programs, and student outcomes.

Domitrovich et al. (2010) and Hamre et al. (2010) took alternative approaches to addressing this question in their examinations of the relationships between implementation variables and student outcomes. Domitrovich et al. included a general measure of teaching quality (i.e., positive climate, sensitivity, behavior management) as a control, whereas Hamre et al. carefully incorporated aspects of general teaching quality into their measure of implementation. Hamre and colleagues found that this measure was related to student outcomes, indicating that overall teacher quality played an important part in implementation quality. Domitrovich et al. found that the teaching quality control variable did diminish the association between implementation and student outcomes, but only on one measure (child aggression).

While these studies indicate that overall teacher quality does impact implementation, Domitrovich et al. (2010) highlight the fact that program implementation was still associated with the majority of student outcomes, even when teacher quality was controlled. In other words, student gains were dependent on effective implementation, rather than teacher quality alone. This has important implications, and offers strong support for the need to approach and measure implementation as a multifaceted construct. This finding may be particularly important in light of the fact that the preschool program in question was rooted in an explicit instruction framework. Domitrovich and colleagues note that critics of explicit instruction in early childhood settings often speculate that explicit instruction in early childhood contexts is unnecessary when quality instruction is in place, but these findings indicate that other program variables, and their effective implementation (e.g., adherence to program theory and active ingredients) clearly impacted student outcomes. Future studies should examine the relationships of these aspects of implementation and other measures of teacher effectiveness.

Program fit. Moving beyond individual teachers' skills to a broader ecological level involves examining the ways teacher, classroom, and school variables align with the fit of specific programs. Each school context is unique, and the alignment between a program and a contextual environment can play a big part in the implementation process (Durlak, 2010; Fixsen et al., 2005; Harn et al., 2013). There are several questions that may indicate whether a program and a school context "match." For instance, does the school have the organizational resources needed to implement the program? Does the program align with other instructional approaches used in the school? Does it match teachers' educational philosophies? Do teachers see the program as important for their

students? Do teachers feel equipped and empowered to adopt and deliver the program? These types of questions highlight contextual variables related to school and teacher buy-in, and while these measures of contextual fit are widely recognized as important for the implementation process, there are few studies that examine how to achieve this fit, or what its actual relationship is with later implementation outcomes (Fixsen et al., 2005).

Fixsen and colleagues (2005) recommend measuring a school's "readiness" for program implementation as a way to assess this contextual fit and teacher buy-in. This includes school-level variables, such as a school's organizational readiness to support program implementation (e.g., training, leadership, resources, work climate) as well as teacher-level variables (e.g., attitudes and beliefs about program, school, professional abilities, student needs). While there is limited empirical evidence about the relationship between these variables and implementation processes, there are some initial studies that indicate that these types of variables are predictive of implementation outcomes. For instance, in a qualitative study of contextual factors that impacted teachers' high or low implementation of an integrated preschool curriculum, Lieber et al. (2009) found that while broader school factors (e.g., training, adult relationships, classroom processes) impacted implementation, teacher attitudes and beliefs (e.g., motivation, belief in program philosophies) were most influential on whether teachers were high or low implementers.

Content and context implications. Collectively, the body of research discussed above suggests that education researchers need to carefully consider *what* to measure in implementation research, and *why*. Table 1 provides an overview of these questions. These considerations should include multiple aspects and dimension of implementation,

contextual variables that may interact with different aspects of implementation, as well as careful attention to program theory, and its relationship with adaptation and active ingredients. This work must also include considerations of when to measure all of these variables, as well as how to measure them. For this, a review of an important methodological practice guide developed by Chomat-Mooney and colleagues (2008) is warranted, as it provides insight into the many measurement issues to be discussed in the remainder of this section.

Chomat-Mooney et al. (2008). In a recent examination of the available tools, measurement approaches, and methodological issues associated with conducting classroom observations, Chomat-Mooney and colleagues (2008) conducted a series of analyses of extant data sets from several large classroom observation studies. These analyses, while focused specifically on measuring classroom quality through direct observation, hold important implications for measuring implementation. Direct observation, while resource-intensive, provides researchers with real time information about the multiple variables that impact program delivery (e.g., program content/active ingredients, teacher quality, classroom or school settings), rather than relying on the student or teacher estimates found in questionnaire or self-report data (Chomat-Mooney et al., 2008; Odom, Hansen et al., 2010; Smolkowski & Gunn, 2012). This flexibility is a tremendous asset when measuring implementation through a multifaceted approach, as researchers can determine what to measure, and train observers to code a wide range of variables accordingly (e.g., multiple aspects of implementation, program content considerations, program context considerations).

Table 1
Research Considerations: What to Measure

Key Questions
Multifaceted Considerations
Which aspects of implementation are important to measure for a given implementation context? o How do these aspects relate to each other? o How do these aspects relate to student outcomes?
Content Considerations
Have the active ingredients been empirically examined and identified? o Which components of the program are critical to achieve the desired effect? o What is the dosage, duration, quality of delivery necessary to achieve that result? Which aspects of implementation reflect the program's theory of change? o Do measures capture more than just adherence to basic procedure? Which components of the program must be adhered to, and which can be adapted to support contextual differences, while still achieving same results? o Which variables (e.g., pacing, dosage, scaffolding, practice opportunities) can be adjusted?
Context Considerations
Where is the school in the implementation process? o Which stage of implementation best defines the current context? How well does the program align or "fit" the school context? o Does the school have the organizational resources to implement the program? o Does the program align with other instructional approaches used in the school? o Does it match teachers' educational philosophies? Are there teacher variables beyond the program that may impact student outcomes? o How do teacher attitudes and beliefs (e.g., motivation, attitudes about professional abilities or student needs, buy-in) impact implementation? o How do general levels of teacher quality impact program implementation?
Key Citations
Dane & Schneider, 1998; Domitrovich et al., 2010; Durlak, 2010; Durlak & DuPre, 2008; Fixsen et al., 2005; Hamre et al., 2010; Harn et al., 2013; Mowbray et al., 2003; O'Donnell, 2008; Webster-Stratton et al., 2011

Specifically, Chomat-Mooney et al. (2008) examined classroom observation data from five studies that collectively included data on over 1900 classrooms, the majority of

which were 3rd or 5th grade classrooms, though one study included a range of 3rd -12th grade classrooms. Observation strategies were fairly consistent across these studies, with each study using global ratings of classroom processes, and/or time sampling of specific operationalized behaviors to measure implementation. Global ratings involve training observers to code classroom processes such as instructional delivery, teacher-student interactions, and classroom management (i.e., instructional implementation) on a Likert-like rating scale to come up with a global measure of implementation quality. Time sampling involves training observers to note the frequency of discrete, observable instructional behaviors in which a teacher or student engages. While Chomat-Mooney and colleagues include a review of multiple classroom observation tools, the studies they included in their analyses used either the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) or the Classroom Observation System-Third Grade (COS-3, NICHD ECCRN, 2002) or -Fifth Grade (COS- 5; NICHD ECCRN, 2004). All three measures are built on the CLASS classroom quality framework, which is based on developmental theory and holds that interactions between teachers and students drive student learning (Chomat-Mooney, 2008). The CLASS is a global measure, whereas the COS-3 and COS-5 include both global and time-sampling measures. Chomat-Mooney et al. (2008) used data from these studies to examine a number of important technical and psychometric issues regarding classroom observation processes. For the purposes of the proposed study, however, it is their examination of the differences between measurement approaches (i.e., time-sampling, global ratings) and how they relate to student outcomes, and timing considerations that are of particular importance.

Relationship between measurement approaches. In their correlational analysis examining the shared variance between different measurement approaches (e.g., time-sampling, global ratings) Chomat-Mooney and colleagues (2008) found few associations between time-sampling observations of instructional time and global measures of classroom quality. In an additional analysis, Chomat-Mooney et al. also examined the relationship between global ratings of classroom quality and time-sampled teacher behaviors to determine the convergence between these two constructs. Interestingly, here again there were very few associations between the two types of measurement approaches, indicating that while these approaches target similar constructs (e.g., quality of instruction, instructional behaviors), these measurement approaches capture different aspects of classroom processes.

Relationship between measurement approaches and outcomes. In addition to examining the relationships between types of observation measures, Chomat-Mooney et al. (2008) also examined how well these measurement approaches predicted student outcomes. They found that these approaches were not only associated with different elements of classroom instruction, but that they were also differentially related to student outcomes. Global ratings of classroom quality were generally related to student outcomes, with higher ratings of quality being associated with lower rates of problem behavior, and higher ratings on specific classroom measures being associated with specific academic gains (e.g., higher rates of productivity and teacher sensitivity were associated with higher reading scores). Some specific time-sampled teacher behaviors (e.g., teaching basic skills, teaching analysis/inference, disciplinary behaviors, attending to student) were predictive of academic and behavioral outcomes, but the strength and

frequency of these associations was less than those found with the global ratings. Collectively, these multiple regression analyses indicate that while both types of observation approaches were predictive of student outcomes, global ratings were more consistently predictive of academic and behavioral outcomes for students than time-sampling measures of teacher behaviors.

Accounting for variance. In an additional set of analyses, Chomat-Mooney et al. (2008) used a multilevel model to examine the sources of variance across the two different observation approaches. Specifically, they constructed their data set to examine observations nested within teachers, teachers nested within raters, and raters nested within school sites. Both time-sampling and global measures had significant rater-level variance, with only 1-7% of the explained variance being attributed to the rater on the time-sampling measure, and 4-14% of the explained variance being attributed to the rater on the global measure. While this smaller observer effect indicates greater reliability for the time-sampling measures, Chomat-Mooney et al. also note that the majority of variance in the time-sampled observations was at the observation level, while the global measures captured much more teacher-level variance. This indicates that *global measures*, while more susceptible to observer effects due to the more subjective nature of the rating scale used in the global measure, were more efficient at capturing *differences in overall teacher quality*. *Time sampling measures*, while more reliable due to the discrete nature of the observation codes, were less sensitive to actual *differences between teachers*, and were better at capturing in depth information of moment-to-moment happenings in the classroom.

Stability over time. Next, using a structural equation modeling framework, Chomat-Mooney et al. (2008) examined the stability of different observation approaches over various timeframes (i.e., school day, school year). Global rating approaches were found to be fairly stable, particularly in terms of capturing constructs related to the emotional and organizational aspects of classrooms. Global measures of instructional supports were slightly less stable, indicating that instructional quality differed according to instructional content and activities. Time-sampling approaches, on the other hand, were much less stable than global ratings, given the sensitivity these measures had to moment-to-moment changes within classroom contexts. Given these findings, Chomat-Mooney and colleagues recommend that researchers conduct at least six observations using time-sampling approaches to reach a stable sample of classroom quality, while they recommend only four observations using a global approach.

Chomat-Mooney et al. (2008) implications. The findings in this report have major implications for the measurement of implementation, particularly in terms of determining measurement approaches. For instance, these findings indicate that global measures may provide more insight into an overall classroom experience (e.g., general teacher/instructional quality), while time-sampling may be more appropriate for identifying moment-to-moment issues in a teacher's classroom or instructional delivery. These findings also suggest that these different approaches should be applied to different parts of the measurement process (e.g., identifying overall implementation quality, examining implementation differences between teachers, addressing specific implementation problems for coaching or training). Based on these results, which need to be replicated, and the findings of differential relationships to student outcomes, Chomat-

Mooney et al. (2008) emphasize the importance of exploring the use and application of multiple observation approaches during implementation research. The next two sections will examine these implications within the broader implementation literature base, and focus specifically on issues related to how and when implementation is measured.

When to measure it. Time is an important contextual factor that needs to be considered during implementation research. Questions around the timing of the implementation process itself are certainly important (e.g., When is a school ready to adopt a new program? When should teachers receive training?) however as Chomat-Mooney et al. (2008) highlight, there are also several time variables that that impact the measurement of implementation (e.g., Frequency across a school year? Time of day?). These considerations will be expanded in the next section to include an examination of two time variables from a measurement perspective: (a) implementation assessment and (b) implementation stage.

Implementation assessment. There is growing recognition of the problematic nature of typical measurement approaches that assess implementation once and use that singular measure to represent implementation across time (Domitrovich et al., 2010; Harn et al., 2013; Mowbray et al., 2003). This approach assumes that implementation measurement is needed only to establish fidelity to treatment to document internal validity (e.g., a singular measure of adherence). Time is but one of these contextual variables, but there is emerging evidence that implementation is not stable and should be assessed multiple times to capture changes across time (e.g., Chomat-Mooney et al., 2008; Domitrovich et al., 2010; Durlak & DuPre, 2008; Fixsen et al., 2005; Zvoch, 2009).

Durlak and DuPre (2008) found evidence that implementation deteriorates over time. They note that only a limited sub-sample of articles examined implementation across time, but based on these findings, they speculate that while early assessments of implementation may be high, program "drift" over time may lead to inflated estimates by the end of implementation. Domitrovich et al. (2010), on the other hand, found evidence that overall levels of implementation of preschool curriculum improved over the course a school year. They suggested that ongoing professional development and coaching support teachers received during the course of the study led to increased teacher understanding and fluency with program materials, which thereby increased implementation. While these are contradictory, they provide strong support for the need to measure implementation across time, and they highlight the importance of considering contextual variables in this measurement process.

In this vein, Zvoch (2009) conducted a multilevel examination of teachers' implementation of early childhood literacy programs and found evidence that implementation improved, declined, or remained stable depending on school context. Specifically, Zvoch found strong site-level differences, with one school demonstrating clear patterns of implementation decline across time, while others demonstrated stable or slightly increasing levels of implementation. Interestingly, however, there were still significant levels of within-site variability, indicating that differences in individual teacher implementation still mattered. Zvoch found that implementation levels were highly variable between teachers at the start of the year, and that differences remained across the course of the year, however school context interacted with these teacher differences to impact implementation across time. For instance, in the school with strong

levels of decline, all teachers' levels of implementation declined, even those who started the year as strong implementers, while in the school with stable moderate levels, the initial differences between teachers remained fairly stable over time (e.g., higher implementers improving over time, lower implementers declining). Clearly, time and context interact to affect teachers' implementation, and implementation measurement must account for these variables.

Multifaceted approaches to implementation assessment. While Zvoch (2009) highlights the impact interactions between time and context have on implementation, it's important to note that this study used a unidimensional measurement approach (e.g., adherence). It is possible that the results obtained are due in part to the co-occurrence of other aspects of implementation. For example, Domitrovich et al. (2010), examined implementation in a multifaceted manner. When looking at their measures of implementation as a whole across the multiple curricular programs, they reported that implementation improved over time. Additional analyses, however, revealed differential patterns for different aspects of implementation. Specifically, Domitrovich and colleagues examined dosage, adherence, and quality. They found that measures of dosage started high and remained so across program delivery. Measures of adherence all began at similarly high levels, but only increased significantly for a more procedurally complex curricular program. Domitrovich et al. speculated that the scripted nature of the other programs helped to maintain high levels of adherence, whereas the more complex activity required increases in fluency and skills that took teachers time to develop.

Domitrovich and colleagues (2010) also measured child engagement and teacher generalization as aspects of implementation quality. Measures of child engagement were

all initially high, but showed patterns of variability over time based on curricular program. Well-developed scripted programs resulted in high levels of engagement across the school year, while more complex or teacher-led programs showed steady increases over the year. One program, which integrated significantly more difficult content over the course of the year showed steady declines. Measures of teacher generalization were initially low, but showed steady growth over the course of the year, indicating that teachers required support, training, and time to improve in this aspect of implementation. Domitrovich and colleagues note that this was consistent with the structure of the study's professional development program, where training sessions differentially focused on generalization over time, with initial sessions targeting adherence and dosage, and later sessions targeting generalization to broader classroom practices.

This professional development structure is also important to note as a broader contextual variable. Teachers in this study received weekly mentoring support in addition to multiple training sessions focused on implementing the various curricular programs under study. This level of coaching and support has important implications for interpreting the overall positive implementation trends found here, particularly when taken in consideration with Zvoch's (2009) findings on the impact of school context on implementation over time. It is also interesting to note that the studies cited in Durlak and DuPre (2008) and Domitrovich et al. (2010) used a mix of teacher report and global observation approaches to measure these various aspects of implementation. Chomat-Mooney et al. (2008) highlighted the fact that the type of measurement approach (i.e., global observation, time-sampling) impacted the stability of measurement across time, as did the time of day in which a measure was taken. Neither of these time variables was

considered in these specific studies, however clearly these additional measurement issues may impact implementation outcomes.

Implementation stage. As Fixsen and colleagues (2005) highlight, implementation is a dynamic process made up of six complex stages that move a program from early adoption into sustainable practice Each stage brings with it unique contextual and implementation challenges that should be considered when developing an implementation measurement approach. For instance, teachers in the early stages of implementation may be grappling with complex issues related to their ability to implement a new program effectively (e.g., acceptance and commitment to new program, alignment with program philosophies, fear of change) (Durlak & DuPre, 2008; Fixsen et al., 2005; Harn et al., 2013). Teachers in the full operation stage, on the other hand, may have worked out some of these initial challenges, and program delivery may more fully reflect stable teacher-level differences in specific program implementation, rather than issues associated with learning to deliver a new program. During innovation and sustainability, teachers may have developed fluency with a program after years of implementation, and begin to make adaptations for a variety of reasons (e.g., program drift, contextual fit, changes in student population), which may both support or detract from student outcomes (Mowbray et al., 2003).

Different stages of implementation are impacted by unique variables, challenges, and implementation processes. Understanding these issues at one stage (e.g., initial) does not necessarily provide insight into the issues impacting implementation at another phase (e.g., innovation). It can also take years for a program to move from the early to latter stages of implementation (Fixsen et al., 2005). Much can change in a school over the

course of a few years (e.g., funding, staff turnover, training needs, changes in student populations), and this time variable highlights the fact that implementation should be thought of as fluid, dynamic, and evolving (Mowbray et al., 2003; Zvoch, 2009).

Implementation measurement should reflect this, yet much of the available research is static, taken at one point in time during exploration or initial stages of implementation (Fixsen et al., 2005). This is problematic on several levels. Measures of implementation from initial stages may be uncharacteristically low, reflecting variables related to teacher learning or the development of systems of support, rather than information about program effectiveness (Fixsen et al., 2005; Harn et al., 2013). The implications from these findings should not necessarily then be used to make decisions about a program's impact on student outcomes, as programs that may in fact be effective may not lead to adequate student growth until they are fully implemented. Measures of these early implementation stages may be better suited to determining what programs look like in a specific context, and providing teachers and schools with support and training as they work toward full implementation (Fixsen et al., 2005). It may also be problematic to generalize findings from measures of early implementation to latter implementation stages. For instance, low measures of implementation during the initial stages may reflect teacher- or school-level issues, indicating that teachers are in need of professional development around specific program components, or that schools lack the specific systems needed to support program delivery. However, as teachers and schools move through implementation stages, those same scores may actually reflect positive adaptions as teachers tailor program features to student and school contexts (Durak &

DuPre, 2008; Harn et al., 2013). Researchers should consider these types of issues related to implementation stage when developing and interpreting implementation measures.

Time variable implications. Researchers not only need to consider what to measure, but also how these multiple aspects of implementation are impacted by the timing and number of measurements, and stage of implementation. The research highlighted above strongly indicates that multiple measures of implementation are needed to develop full understanding of program delivery processes, and to determine how implementation variables relate to student outcomes. Table 2 provides an overview of these issues. Developing this type of multifaceted implementation measurement approach across time has huge implications for the field, however, as Flay et al. (2005) indicate, much of this work is not yet common practice, which may, at least in part be due to the fact that there are still several methodological concerns related specifically to measurement processes. Toward this purpose, the next section will explore some important methodological issues that impact how researchers go about measuring the multicomponent implementation variables.

How to measure it. Of all of the complex and important issues involved in considering implementation as measurement in a context, perhaps most important from a measurement perspective is thinking about *how* to measure it. Yet without accurate, valid, and efficient measures, much of the discussion around these complexities (e.g., content considerations, contextual variables, time considerations) remains speculative. This refrain is consistently noted in the implementation literature (e.g., Durlak, 2010; Durlak & DuPre, 2008; Fixsen et al., 2005; O'Donnell, 2008; Van Meter & Van Horn, 1975), though technological advances in observation (e.g., video, electronic coding

Table 2
Research Considerations: When to Measure It

Key Questions
Time of Assessment Considerations
Is implementation expected to increase, decrease, or remain stable based on program theory and context of delivery? o How many times should implementation be assessed across program delivery? o How many times per year? Across multiple years? Which aspects of implementation should be measured at which points in time? What type of measurement approach should be used across time? How do measurement decisions account for the interaction of time and context? o Do measures of implementation account for expected changes across time (e.g., pre/post measures that align with PD or coaching sessions)?
Stage of Implementation Considerations
Which stage of implementation best describes the school's current context? o What is the purpose of implementation measures for that given stage (e.g., to inform PD, to identify system-level issues, to evaluate outcomes)? o Which aspects of implementation should be measured at each stage? o Are there implementation benchmarks or thresholds that should be set as goals for a given stage?
Key Citations
Chomat-Mooney et al., 2008; Domitrovich et al., 2010; Durlak & DuPre, 2008; Fixsen et al., 2005; Mowbray et al., 2003; Zvoch, 2009

systems) and analysis (e.g., multilevel modeling) allow for more in depth measurement that capture some of these complexities (Chomat-Mooney et al., 2008; Domitrovich et al., 2010; Mowbray et al., 2003; Zvoch, 2009). As such, the final measurement consideration in this literature review will focus on critical types of measures and measurement approaches that can support researchers in understanding implementation and its relationship to student outcomes.

Measurement approach. Implementation is typically measured using one of four commonly used measurement approaches: (a) direct observation, (b) self-report, (c) interview, and (d) archival records (Chomat-Mooney et al., 2008). While each of these

brings distinct advantages and disadvantages to the measurement process, direct observation, with its ability to document real-time interactions about the multiple complex variables that impact program delivery, provides researchers with more accurate, in depth views of implementation process and the variables that impact student outcomes. As such, these direct observations can provide information that can be used to improve classrooms, teachers, *and* student outcomes via multiple mechanisms (e.g., environmental improvements, curricular changes, professional development, coaching) (Chomat-Mooney et al., 2008; Harn et al., 2013; Pianta & Hamre, 2009; Smolkowski & Gunn, 2012).

The information that is obtained during classroom observations depends on the theoretical framework, perspective, and observation approach that is used to obtain the information. Multiple aspects of classroom interactions can be observed, such as observations that target the overall classroom environment, general classroom or teacher quality, content-specific interactions, or teacher-student interactions. These interactions can also be assessed within various time frames, from a broad perspective on global features of classrooms across time, to discrete assessments of classroom processes in momentary time samples. Classroom observations can also be measured in various ways. Qualitative approaches can be used to provide rich descriptions of individual classrooms, teacher and student attitudes and intentions through the use of various methods such as ethnographic fieldwork, critical analysis, or interviews. Quantitative approaches can be used to provide clear measurement of specific, operationally defined behaviors or classroom variables, such as the number of instructional units delivered (e.g., number of items covered) or the frequency of specific teacher or student behaviors. Deciding which

of these elements to include in a classroom observation is an empirical question. Issues such as construct validity, reliability, and technical aspects of observational approaches (e.g., rater effects on observations, stability of indicators) must all be considered, as various observational approaches address these issues in different ways (Chomat-Mooney et al., 2008). This final section will expand on the two direct observation approaches discussed in the Chomat-Mooney et al. (2008) report: time-sampling, which is a type of discrete behavioral measurement approach, and global ratings of implementation processes. Both of these observation approaches are quantitative; they use standardized observation approaches to quantify implementation processes, and then link those measurements to student outcomes. Both of these types of observation approaches have successfully been used to develop various standardized measures that have been found to have adequate reliability and validity, that demonstrate strong patterns of generalizability, and that have the potential to be used at-scale (Chomat-Mooney et al., 2008). Despite these similarities, behavioral and global rating approaches have very different theoretical assumptions that impact the way these approaches are used to measure implementation and how they relate to outcomes.

Discrete behavioral observation approaches. Discrete behavioral observation approaches are rooted in a behavioral framework, and aim to quantify important behavioral events in consistent, transparent ways that allow researchers to document and evaluate patterns in behavior across time. Behaviors of interest must therefore be directly observable, able to be operationalized into discrete, easily understood and agreed-upon terms (i.e., occurring, non-occurring binaries), which can be quantified, and as such measured accurately (Kennedy, 2005). A major advantage to this type of measurement

approach is that objectivity is maximized. Only those behaviors that can be observed and that meet the clearly-defined decision rules in an operational definition are measured, and issues related to observer judgment and inference that can impact reliability are reduced (Kennedy, 2005; Smolkowski & Gunn, 2012).

Behavioral approaches to direct observation require rigorous measurement procedures (see Kennedy, 2005 for an in-depth overview). When developing this approach, researchers need to consider several components of their measurement system. For the purposes of this study, two key measurement considerations related to discrete behavioral observation will be addressed: dimensional qualities and behavior sampling. Dimensional qualities refer to different characteristics of a behavior in space and time, such as frequency or duration, which can be quantified through direct observation (Kennedy, 2005, p. 82). Behavior sampling refers to the strategies that are used to collect data about behaviors across time, such as event, partial-interval, whole-interval, and momentary-interval recording. Each of these strategies has different advantages and disadvantages, however the key considerations are whether a behavior needs to be counted each and every time it occurs (i.e., event recording), or if some sort of time-sampling strategy (e.g., partial-interval recording) can be used to estimate how often the behavior occurs over time. Researchers using time-sampling identify set intervals (e.g., 15-sec) during which observers score whether or not a behavior occurs, though the different sampling techniques have different requirements for whether a behavior has to occur for the entire interval (i.e., whole-interval recording), at any time during an interval (i.e., partial-interval recording), or only during a specified part of the interval (i.e., momentary-interval recording) (Kennedy, 2005). These measurement decisions are

highly contextual, and depend on the target behaviors of interest, and the settings in which researchers are observing.

Chomat-Mooney et al. (2008) noted that behavioral approaches, such as time-sampling, in the context of instructional implementation involve the direct observation of discrete teacher and student behaviors, interactions, and chains of behavior (i.e., antecedent and consequent events around a target behavior) that impact program delivery and student learning. While behavioral measurement approaches are highly dependent upon research contexts and individual student behaviors, there has been a fair amount of consistency in how these approaches are applied when examining instructional implementation.

For example, in a study examining academic instruction for students with emotional and behavioral disorders (EBD), Sutherland, Alder, and Gunter (2003) examined several important teacher and student target behaviors. For teachers, they examined how many opportunities to respond (OTRs) they provided students, as well as teachers' use of praise. For students, they examined correct academic responses (CAR), problem behaviors (PB), and on-task behavior. The measurement system was tailored to each behavior (i.e., frequency, event recording for OTRs, praise, CARs, PB; frequency, momentary-interval for on-task behavior), based on considerations of each behavior's dimensions, frequency, and observational issues. In another study examining professional development strategies for supporting teachers' instructional practices, Stichter, Lewis, Richter, Johnson, and Bradley (2006) took a similar measurement approach, examining OTRs, teacher feedback, and student PB, on-task and off-task behavior, in addition to several indirect measures (e.g., student social skills, archival records of academic

performance). Here again the measurement system was individualized, with a mix of frequency and duration measures that were assessed using momentary-interval recording.

With their tool The Classroom Observation of Student-Teacher Interactions (COSTI), Smolkowski and Gunn (2012) built on these types of measurement systems, and extended their application to look at important fine-grained teacher-student interactions during early reading instruction in general education kindergarten classrooms. The COSTI examines four interactions: (a) teacher demonstrations, (b) student independent practice, which is similar to OTRs, (c) student errors, and (d) corrective feedback. These interactions are coded using serial event coding, where each occurrence of the target behaviors are recorded in the sequence in which they occur. This involves documenting both the frequency of individual target behaviors, and the duration of behavioral chains, or teacher-student interactions, which can then be converted to rate per minute.

The target behaviors and teacher-student interactions observed using these time-sampling approaches have strong support (e.g., Archer & Hughes, 2011; Brophy & Good, 1986; Rosenshine, 1997), and the types of observation studies cited above underscore their importance as well. Sutherland et al. (2003) found that increases in teacher OTRs led to increases in student CARs and on-task behavior. Stichter et al. (2006) found that while increased OTRs did not lead to changes in directly observed student behaviors, it did lead to increases in students' literacy scores. Smolkowski and Gunn (2012) found that students' independent practice, which is strongly related to OTRs, was a consistent predictor of students' reading outcomes. Collectively, these studies represent strong

examples of the ways behavioral observation approaches can be used to measure aspects of instructional implementation that clearly impact student outcomes.

Global ratings. Global rating observation approaches are rooted in quantitative measurement methodologies, and aim to examine broad features of settings in an effort to measure the mechanisms, experiences, and processes that impact outcomes on a large scale (Chomat-Mooney et al., 2008; Hamre, Pianta, Mashburn, & Downer, 2007; Pianta & Hamre, 2009). These broader features can be related to several observable constructs, such as the physical characteristics of a setting, specific behaviors of individuals in that setting, or interactions between individuals in that setting. The outcomes of interest, setting, and theory dictate what is observed, and those constructs are operationalized into broad dimensions that thoroughly describe the construct of interest (e.g., classroom management, instructional support), usually through a rubric or measure guide. Global rating approaches use rating scales, whereby observers code the degree of occurrence for each dimension (e.g., strongly present, present, weak, absent), rather than a dichotomous rating of occurrence or non-occurrence. Anchors descriptions outline the varying degrees of each dimension, and observers then decide where each observed setting falls along that spectrum (Fitzpatrick, Sanders, & Worthen, 2011; Hamre et al., 2007).

This process allows for more nuance and variability than a dichotomous measurement approach, as observers are able to make judgments about the quality of the settings they observe, but it also introduces more observer subjectivity into the measurement process (e.g., Smolkowski & Gunn, 2012; Zvoch, 2009). Global rating observation approaches often require more in-depth training than more objective measurement approaches, as observers may need to develop more understanding of an

observed program's theory, and the aspects that distinguish one rating degree from another. While some concerns persist about the objectivity of such measurement approaches, guided practice and reliability checks are often used to ensure that observers reach specified reliability criteria, and many global rating measures have been documented to have strong reliability, and be highly predictive of program outcomes (Chomat-Mooney et al., 2008; Pianta & Hamre, 2009).

Standardized global rating approaches have been used throughout the literature to examine and document a number of broad, theoretically important constructs at the classroom level that impact student learning. The majority of this work has been done in the early childhood field, and there are several measures that have been developed that have strong psychometric properties (Chomat-Mooney et al., 2008; Hamre et al., 2007; Pianta & Hamre, 2009). For example, Mashburn et al. (2008) examined three different types of global measures of preschool quality (i.e., the NIEER, an infrastructure metric, the ECERS-R, a broad classroom environment metric; and the CLASS, a teacher-student interactions metric) and their associations with a range of student outcomes. While the NIEER used a self-report approach (i.e., teachers or administrators provided information about school or classroom infrastructure), the ECERS-R and CLASS involved classroom observations across a range of dimensions. With both of these measures, observers rate classroom quality on a scale of 1 to 7, though each tool has different anchors (e.g., 1 = inadequate quality vs. 1-2 = low quality). Observers were trained to reliability using videotapes or practice classroom observations prior to data collection for both of these tools. In another study, Maxwell, McWilliam, Hemmeter, Ault, and Schuster (2001) used a global observation approach to examine developmentally appropriate teacher practices

in K-3 classrooms. Here again the global observation measure targeted a range of broad classroom constructs, as opposed to instructionally specific content (e.g., teacher-child language, instructional methods, appropriate transitions, children's role in decision-making), and each item was rated on a 7-point scale, (e.g., 1 = developmentally inappropriate practice). This measure includes a sub-rating (i.e., yes, no, N/A) for each anchor point, which may provide an additional observer check and reduce subjectivity. As is common practice, observers were trained to ensure that coders were within one point of other coders on 80% of items. Both Mashburn et al. (2008) and Maxwell et al. (2001) documented that the global measures used in these studies represented reliable, internally consistent, psychometrically valid tools. Mashburn et al. found, however, that of the three observation tools they examined, only the CLASS was strongly predictive of student academic, language, and social outcomes. Maxwell et al. found that their global measure predicted teacher practices, however they did not extend this to examine whether or not these practices resulted in improved student outcomes.

Researcher considerations. Collectively, these examples of discrete behavioral observation and global rating approaches, and the analyses examined in the Chomat-Mooney et al. (2008) report highlight both the advantages and disadvantages of these two measurement approaches. In the Chomat-Mooney et al. report, while both approaches had adequate reliability, time-sampling was more reliable than global measures, which may have important implications when using these approaches to measure specific aspects of implementation. Chomat-Mooney et al. also reported differences in the way these approaches account for variability, with time-sampling accounting for more moment-to-moment variability, and global ratings accounting for more teacher or classroom level

differences. Stoolmiller, Eddy, and Reid (2000) discuss the fact that time-sampling measurement approaches are particularly useful for capturing information about behaviors that are "state dependent," rather than trait-like. This difference draws from psychological theories that stable personality, temperament, or behavioral characteristics are traits, whereas characteristics that vary markedly across contexts are states (Stoolmiller et al., 2000). In the context of classroom observations, some aspects of classrooms (e.g., teaching style) may be quite stable and trait-like, while other aspects (e.g., specific practices) may be quite variable, and would be better conceptualized as state-like (Cobb & Smith, 2008; Klingner, Boardman, & McMaster, 2013). From a measurement perspective, then, this indicates that different aspects of instructional implementation may align differentially with different types of measurement approaches.

Measurement approach implications. While both discrete behavioral observation and global measurement approaches have solid empirical support for their use in educational research, there is still limited research comparing and contrasting these approaches in implementation research. Chomat-Mooney et al. (2008) provide an excellent example of the lines of research that should be examined, however this work should be expanded to examine different types of measures, different student outcomes, and the possibilities of using multifaceted, multidimensional measures that integrate these two measurement approaches. Understanding the ways these measurement approaches map on to specific aspects of implementation has important implications for researchers, and taking a multifaceted approach that incorporates both time-sampling and global ratings may provide more information about implementation, and how it impacts outcomes. These questions about how to measure implementation should be considered

as a final, and vital piece of the puzzle involved in approaching implementation research as measurement in a context. Table 3 provides an overview of these questions.

Outcomes

Typical approaches to implementation measurement often do not document the relationship between implementation and outcomes. Questions also remain about how implementation relates to different types of student outcomes (e.g., academic, social), and whether implementation is differentially important for different subgroups of students (Durlak & DuPre, 2008; Harn et al., 2013). Despite these limitations, when examined collectively across the broader human services field, there is strong support for the fact that better implementation is associated with better outcomes (Durlak & DuPre, 2008; Fixsen et al., 2005; O'Donnell, 2008). This section will summarize the few studies that have linked implementation to outcomes.

Measurement considerations. "Better" implementation, however, all depends on what you measure, and when and how you measure it. Relating implementation to student outcomes is an equally complex process, yet the majority of research that has undertaken this task is still rooted in typical measurement approaches. This may include requiring that teachers reach unrealistic or rigid levels of implementation (e.g., 100% adherence) or using unidimensional measures. Each of these measurement considerations has implications for how implementation variables are used to interpret student outcomes.

***Levels of implementation*:** One major issue involved in relating implementation outcomes to student outcomes is that there is little agreed-upon understanding of the levels of implementation that are required to achieve positive student outcomes (Durlak

Table 3
Research Considerations: How to Measure It

Key Questions
Measurement Approach Considerations
What type of direct observation approach (i.e., behavioral, global rating) will accurately and reliably capture implementation information? o What behaviors or constructs are important to observe? o How will observers be trained to reliability? o What resources are necessary to support the use of this measurement approach? Is there a specific level of variability (i.e., classroom or teacher-level differences, moment-to-moment differences) that is theoretically important for implementation of a given program? o Which observation approach is more likely to capture those types of differences? How does each direct observation approach align with the different aspects of implementation? o Should different approaches be used to measure different aspects of implementation? o Should a multicomponent measure be used?
Key Citations
Chomat-Mooney et al., 2008; Hamre et al., 2007; Pianta & Hamre, 2009; Smolkowski & Gunn, 2012; Stoolmiller et al., 2000

& DuPre, 2008; Harn et al., 2013). Higher levels of implementation are generally associated with higher outcomes, however there is little empirical evidence to suggest that the levels of implementation often required during efficacy studies (e.g., 85% or higher) are necessary to achieve positive outcomes. Durlak and DuPre (2008), in their discussion of threshold effects, cite evidence that the studies with implementation levels as low as 60% obtained positive results, and they note the strong positive associations between program outcomes and implementation levels that were well below 80% for the majority of programs in their study. Mowbray et al. (2003) emphasize the fact that the nature of the program may impact implementation thresholds, noting that high levels of adherence (i.e., near perfect) may be more appropriate with standardized, well-specified

programs (e.g., scripted programs), but that other less-structured programs may be more suitable to adaptations (i.e., adaptations that don't contradict program theory) that support students in a given context, and lower levels of implementation may therefore be acceptable. These questions need to be studied systematically and examined in terms of maximizing student outcomes before the relationship between levels of implementation and student achievement can be fully understood (Durlak, 2010; Durlak & DuPre, 2008; Harn et al., 2013).

Multifaceted approaches. While studies that examine multiple measures of singular aspects of implementation offer important insights to the field, there is increasing recognition that a multidimensional approach to implementation measurement (e.g., multiple aspects, structural and process dimensions) is necessary to truly begin to untangle the complex relationship between implementation and intervention effectiveness and student response to intervention (Harn et al., 2013; Odom et al., 2010; O'Donnell, 2008). While limited, several studies in this research vein have found evidence that different aspects or dimensions of implementation may be differentially related to student outcomes. For instance, Odom et al. (2010) examined the relationships of structural (dosage), process (quality), and multicomponent (composite) measures of implementation to a variety of preschool outcomes. Both the structural and multicomponent measures were found to be associated with math outcomes, while the process and multicomponent measures were associated with social behavior outcomes. Only the process measure was associated with literacy outcomes.

Crawford, Carpenter, Wilson, Schmeister, and McDonald (2012) examined the relationship between implementation variables and student outcomes in a computer-based

middle school math program. Crawford et al. found strong associations between student outcomes and their three structural variables, which included two measures of dosage and an adherence composite that measured both teacher adherence to program and student engagement. Adherence, in particular, had such a strong effect that even slight decreases required a significant increase in program dosage to maintain student outcomes. Their process composite, which used a rating scale to measure a range of teacher behaviors (e.g., classroom management, data-based decision making, technological problem-solving, communication with researchers, time management), was not significantly associated with student math outcomes.

Domitrovich et al. (2010) examined dosage, adherence, and two measures of quality (child engagement and teacher generalization), in their study of the implementation of a comprehensive curriculum in Head Start. Dosage and generalization were not associated with student outcomes, though the authors note that the small sample size (n = 22) and limited variability across these measures may have been factors. Adherence and child engagement were highly related to social behavior outcomes, though no positive associations were found with literacy outcomes, and in fact adherence and generalization were negatively associated with two literacy measures. Domitrovich et al. posit that this may be due more to a lack of fit between these measures and the active core ingredients of the interventions than an actual negative relationship between the interventions and literacy outcomes.

In their study of a supplemental literacy and language development curriculum, Hamre et al. (2010) examined the effects of multiple measures of dosage, adherence, and quality aspects of implementation on student outcomes. Adherence was not associated

with any student outcomes, while one measure of dosage was positively associated with preschool students' emerging literacy scores. One measure of quality of delivery, teachers' classroom literacy focus, was positively associated with print awareness and emerging literacy scores. Hamre et al. also found an interaction between child-level characteristics and their two measures of implementation quality. Students with weaker emerging literacy pre-test scores and students who spoke a language other than English at home both had significantly higher growth on literacy outcomes when quality aspects of implementation were higher in their classroom.

Outcomes implications. Albeit small, this literature base collectively highlights the fact that different components of implementation are differentially related to various types of student outcomes. Fixsen and colleagues (2005) commence their review of the implementation literature base by highlighting the difference between implementation outcomes (e.g., Is the program being delivered as intended? By whom, to whom, in what contexts?) and program/student effectiveness outcomes (e.g., Does the program, when implemented as intended, result in improved student learning?). This caution is important, particularly in terms of considering the role of implementation research in understanding the variables that impact long-term sustainability and improvements in student learning. As implementation is considered to be a mediator or moderator of program effectiveness, examining implementation outcomes in and of themselves is an incredibly important process (Mowbray et al., 2003), however this work should be undertaken toward the purpose of informing the interpretation of student outcomes. Understanding how implementation varies across contexts can provide valuable information to researchers as they interpret student outcomes (e.g., Should low outcomes

be attributed to low levels of implementation? Were positive outcomes achieved with varying levels of implementation? Which aspects and program active ingredients were consistent across settings? Were there threshold levels that must be adhered to in order to achieve program effects?) (Flay et al., 2005; O'Donnell, 2008).

A multifaceted approach to implementation measurement is critical to this process, as it not only highlights aspects of implementation that can be used to support the implementation process more efficiently (e.g., supporting teachers through coaching), but also can provide researchers with much more targeted information about the aspects of implementation that matter for ensuring that a given program has an effect for its intended participants (Durlak & DuPre, 2008). While this work is a complex and challenging process, it is absolutely critical for addressing the research to practice gap that persists due to a lack of understanding of the relationships between implementation and intervention outcomes. The proposed study is an attempt to contribute to that process, and aims to answer some of the many questions outlined throughout this chapter.

The Current Study

Implementation is a complex task. Implementation measurement must therefore be an equally nuanced process to account for these complexities. Researchers must consider which aspects of implementation should be measured in a given context and situation. This involves examining program theory, active ingredients, and contextual factors. In addition, they must consider how to time the measurement process, and how time across the implementation stages impacts program delivery, and measurement. Researchers must also consider which measurement approaches fit a given context, and

the specific implementation research questions that are being targeted. Collectively, this involves decisions about not only what to measure, but when to measure it, and how.

Each of these is a complex task in and of itself, and warrants empirical examination. Understanding how all of these complex variables relate to each other, and to student outcomes is largely an unknown in the field. The current study is an attempt to address but one part of the implementation measurement puzzle: understanding how different measurement approaches relate to each other and student outcomes when holding the context constant. Specifically, this study examined three different measurement approaches in their relationships to outcomes in schools sustaining the use of EBPs. Toward this purpose, only the measurement approaches were systematically varied, while the instructional content (i.e., what to measure) and time (i.e., when to measure it) variables remained consistent through the use of videotaped instruction. The specific research considerations made about (a) what to measure, (b) when to measure it, and (c) how to measure it will therefore be described in detail below. The measurement context, which will be described in depth in the next chapter, will therefore be briefly overviewed to situate these research decisions and align them with the literature review.

Research context. The current study examines the implementation of EBP reading instruction, delivered to kindergarten students at-risk for reading failure. Instructional aides (IAs) deliver these interventions as part of a school district's established RTI framework, however the focus of this study is classroom-level processes, rather than broader school factors. All IAs in the study have years of experience with the specific EBPs they are delivering, all of which are based on theories of explicit instruction. Theoretically, an explicit instruction framework is based on principles that

integrate effective content delivery, instructional and classroom organization, and maximized student engagement (Archer & Hughes, 2011). Within this framework of explicit instruction, there are several elements of content delivery, (e.g., the use of clear, consistent language, brisk pacing, and clear modeling of skills and strategies prior to student practice) that have been associated with increases in student achievement across both academic and social content areas (e.g., Chard, Vaughn, & Tyler, 2002; Kroesbergen & Van Luit, 2003; Simonsen et al., 2008; Swanson & Hoskyn, 1998; Vaughn, Gersten, & Chard, 2000). Instructional and classroom organizational elements (e.g., teachers' familiarity with lessons, the use of targeted grouping strategies, time management that maximizes guided and independent practice) are also linked to increases in student learning. Instructional delivery also depends upon student responsiveness, so elements such as frequent monitoring of student performance, promoting active engagement, and delivering students with immediate corrective or positive feedback are also critical for maximizing student learning (Archer & Hughes, 2011).

What to measure. Decisions about what to measure were based on careful considerations of the instructional content and context under review. All eight aspects of implementation could have potentially highlighted important variables about instructional implementation in the current study, however resource (e.g., issues of efficiency, cost, methodological parsimony) and theoretical considerations (e.g., curriculum-independent delivery, full implementation stage) led to the selection of four aspects of implementation: (a) adherence to program theory, (b) dosage, (c) quality, and (d) participant responsiveness. These aspects were selected primarily because each

represents an aspect of implementation that requires interaction between teachers, students, or materials, which aligns with the explicit instruction framework of the specified EBPs. They can also be assessed independently from a specific curricular program, unlike typical measures of adherence to program procedure and program reach. These aspects of implementation also represent both structural (adherence, dosage) and process (quality) dimensions of implementation, and participant responsiveness includes both dimensions, depending on how it is measured. Collectively, these aspects were theorized to provide important insight into the ways these specific subcomponents of implementation impact the delivery of targeted small group instruction, and how these variables are related to outcomes for at-risk kindergartners.

Each of the three direct observation tools being examined in the study was used to represent different components of implementation. The OTR approach targeted structural aspects of instructional implementation (i.e., dosage of teacher instructional behaviors). The CLASS targeted process-oriented aspects of implementation (quality of delivery, participant responsiveness). The QIDR is an integrated tool, and was used to examine both structural and process-oriented aspects of implementation (adherence to explicit instruction program theory, quality of delivery, multiple aspects of student responsiveness).

When to measure it. As with what to measure, decisions about when to measure implementation were made with careful consideration of the research and implementation context. Implementation was assessed, through the use of video, on a weekly basis throughout the entire intervention program. These weekly videos were used to examine both average implementation across groups, and changes in implementation over time.

The small group reading interventions under study were delivered as part of a school district's established RTI framework (i.e., over 10 years of implementation), and as such these programs were considered as occurring during full implementation. Given this implementation stage, the measurement focus was no longer on evaluating whether the programs work, but rather on examining the factors that influence how well teachers deliver the program, and how this impacts student response.

How to measure it. The purpose of the current study was to compare commonly used classroom observation measurement approaches (i.e., discrete behavioral observation, global rating); as such decisions about how to measure implementation revolve around this comparison. As this study uses video, each of these approaches was used to assess the *same* instructional implementation, which allowed all other implementation variables to be held constant at minimal observation cost. Specifically, this study examined one tool that used a discrete behavioral observation approach (OTR), one that used a global rating approach (CLASS), and one tool that used an integrated approach (QIDR) to observe the same instructional implementation, and then to compare how these measures are related to each other, and to student outcomes.

These tools were specifically selected to provide insight into whether moment-to-moment differences (i.e., discrete behavioral observations), teacher- or group-level differences (i.e., global ratings of instruction), or an integrated approach were more strongly associated with student outcomes in the current research context. The multifaceted approach being used here, while not comprehensive, was also selected to provide insight into the ways these varied measurement approaches map on to the specified aspects of implementation, particularly in the context of examining the fully

operational implementation of small group explicit instruction. Table 4 presents an overview of both *what* is measured, and *how,* by each of the three direct observation tools, which are described in detail in the next chapter.

Table 4
Overview of Measures

Tool Components	Key Measures		
	OTR	CLASS	QIDR
Implementation Subcomponents			
Structural Dimensions			
Adherence			X
Dosage	X		
Responsiveness			X
Process Dimensions			
Quality		X	X
Responsiveness		X	X
Measurement Approaches			
Behavioral Observation	X		X
Global Rating		X	X

Note. OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt.

Research Questions

Based on the conception of the specific measurement context outlined above, the current study addressed the following research questions:

1. How do three implementation/observation tools relate to each other?

2. How do the observational tools relate to student outcomes? Which observational tool individually accounts for the most variance in student outcomes?
3. How do the observational tools uniquely account for variance in student outcomes when entered into the model simultaneously?
4. What does implementation look like across intervention time by implementation tool?

CHAPTER III

METHOD

> "Although interest in both the practical issues of implementation and implementation research have grown dramatically in the past few years, a fundamental issue that warrants further understanding is the development of measures that are both valid and reliable to assess implementation quality." (Greenberg et al., 2005; p. 56)

Given this study's focus on comparing three observational measures of instructional implementation, and their relationships to student outcomes, this study examined an extant data set comprised of 64 videos of small group instruction with at-risk kindergartners. This ensured that multiple elements of instructional implementation (i.e., delivery, context, time) could be held constant, so that the observational approaches used could be examined to see how these tools relate to this specific implementation context, and to student outcomes. As such, the current chapter provides (a) an overview of the extant data set and the intervention procedures, (b) a description of the specific observation tools, and (c), the specific procedures used to evaluate implementation.

Participants

Settings. This study analyzed data that were collected from two elementary schools in a school district of 11 schools, with approximately 5,700 students, in a mid-size city in the Pacific Northwest. School 1 had 680 students in kindergarten through eighth grade enrolled during the 2011-2012 school year. Of these students, 75% were Caucasian, 17% were Hispanic, and the remaining 8% were Asian/Pacific Islander (3%), Black, not Hispanic (2%), American Indian/Alaskan Native (2%), or from an unspecified racial/ethnic background (1%). Fifty-one percent of students at School 1 were eligible for free or reduced-price lunch. School 2 had 339 students in kindergarten through fifth grade enrolled during the 2011-2012 school year. At School 2, 66% of students were

Caucasian, 26% were Hispanic, and the remaining 8% were Black, not Hispanic (4%), American Indian/Alaskan Native (2%), or Asian/Pacific Islander (2%). Of these students, 64% were eligible for free or reduced-price lunch. Students within our project were similar in ethnic distribution to the school demographics and 32% of the students were identified as English language learners and 21% were receiving special education service.

Interventionists. Seven instructional assistants delivered the Super K small group instruction during the intervention. The interventionists had an average of 11 years teaching (9-15), and an average of 9.8 years (3 – 14) using the specific reading programs used in this study. All interventionists were women; four were Caucasian, two were Latina/Hispanic, and one was multiracial. Four interventionists held high school diplomas, and three had their associate's degree. Table 5 presents an overview of study interventionists by school. Instructional assistants were the typical intervention providers in the school and consented to participating in the project.

Student participants. Students were selected using school-determined screening approaches for identifying students at-risk for reading difficulties (see screening procedures section below). Using this process, 37 students were identified across the two schools. Consent was obtained for 35 students, however 4 students moved prior to the end of intervention, leaving a final sample of 31 students. Of the 31 students, eleven were classified as English language learners (ELLs), and seven were receiving special education services. Table 6 presents an overview of student demographic information by school.

Table 5
Interventionist Characteristics

Title	Race/Ethnicity	Highest Degree Earned	Years Teaching	Years at School	Years using Intervention(s)
School 1					
I.A.	Caucasian/White	HS	15	15	12
I.A.	Caucasian/White and Hispanic	A/A	12	12	12
I.A.	Latino/Hispanic	HS	9	9	3
I.A.	Latino/Hispanic	HS	14	6	14
School 2					
I.A.	Caucasian/White	A/A	9.5	9.5	9.5
I.A.	Caucasian/White	A/A	10	10	10
I.A.	Caucasian/White	HS	9	9	9

Note. I.A. = Instructional Assistant; A/A = associate's degree; HS = high school.

Table 6
Student Demographic Information

	N	Boys	Girls	Identified ELL	Active IEP
School 1	15	9	6	5	4
School 2	19	13	6	6	3

Intervention Measures and Procedures

Student measures. To identify students in need of intervention, the Dynamic Indicators of Basic Early Literacy Skills *(DIBELS;* Good & Kaminski, 2002) was used to assess students' early literacy skills including measures of letter recognition, phonological awareness, alphabetic principle, and fluency. In this study, the DIBELS Letter Naming Fluency (LNF) and Initial Sounds Fluency (ISF) were administered to students in the fall. Each measure is a brief, individually administered, standardized assessment. During LNF administration, students are asked to name as many letters as they can in one minute when presented with a page of random upper and lowercase

letters. The one-month, alternate form reliability for LNF is .88 in kindergarten. The predictive validity of kindergarten LNF with first grade Woodcock-Johnson Psycho-Educational Battery-Revised Reading Cluster standard score is .65. ISF assesses a child's ability to identify a picture that begins with a specific sound as well as produce the initial sound(s) of an auditorially presented word. The alternate-form reliability of the ISF measure is .72 in January of kindergarten and it predicts .36 to the Woodcock-Johnson Psycho-Educational Battery total Reading Cluster score.

Screening procedures. Students were selected to participate in the Super K study based on the results of district screening procedures obtained in the fall of the 2011-2012 school year. All kindergarten students were screened with the LNF and ISF measures. Students most at-risk (i.e., ISF< 3 and LNF< 3) were invited to participate in the project.

Dependent measure. The Word Attack (WAT) subtest of the Woodcock Reading Mastery Tests—Revised (WRMT-R; Woodcock, 1987) was the final student outcome variable for this project. The WRMT-R is a standardized, individually administered, non-timed measure of essential reading skills. The WAT subtest measures a student's ability to decode nonsense words of increasing difficulty. A median split-half reliability coefficient of .87-.94 is reported for the WAT subtest in the standard sample. Concurrent validity ranges for the subtests of the WRMT-R are reported to be from .63-.82 when compared to the Total Reading Score of the Woodcock-Johnson Psycho-Educational Battery (Woodcock & Johnson, 1977). This measure was administered at the beginning and end of the intervention.

Intervention procedures. All students who participated in the Super K programs received supplemental reading instruction using evidence-based programs for

approximately 30 minutes daily. This intervention occurred outside of typical instructional time (i.e., before or after school) and was in addition to instruction using the school's core reading program. Students received instruction in small groups of three to five for an average of 34 instructional days (range 28-41). Both schools used either Reading Mastery (Engelmann et al., 2002), Early Reading Intervention (*ERI;* Simmons & Kame'enui, 2003), or a combination of these two curricular programs to provide systematic, explicit instruction aimed at increasing students' early reading skills. Both programs are scripted, supplemental programs that use direct instruction principles, including teacher modeling, frequent opportunities to respond, highly scaffolded practice opportunities, and corrective feedback to systematically develop students' phonological awareness and alphabetic principle skills (Carnine, Silbert, Kame'enui, & Tarver, 2009).

Implementation Data

The purpose of the current study was to examine the relationships of three direct observation tools (OTR, CLASS, QIDR), and their associations with student performance on WAT in this kindergarten data set. While the actual instructional data were collected through the use of video as part of a larger study, the focus of this study was the collection of implementation data using this data set. As such the next section will describe the data set, as well as the implementation measures and data collection procedures in depth.

Video data set. As part of the Super K program, the seven interventionists completed weekly video recordings of intervention implementation across the 7 weeks of the study. One interventionist worked with two groups, such that videos were obtained across eight small groups. Research assistants delivered video cameras to participating

IAs weekly. Each school generally designated a specific day of the week to tape instructions (e.g., School 1 taped lessons on Tuesdays), unless a scheduling issue occurred. At the start of each lesson, the IA would place the video camera strategically so that all students and IA instruction could be seen and heard. At the end of each week, research assistants would collect the cameras and download the videos onto a secure database. A range of seven to nine videos were collected from each group, leading to a video data set of 64 videos. Videos were approximately 25 minutes in length (range 14:58-30:50).

Implementation measures. The use of videos allows for repeated measurement of implementation on each observation implementation tool across time, controlling for interventionists and students, thereby highlighting the differences between the three implementation tools. These tools all examine instructional implementation through a focus on teacher-student interactions, but vary in their approaches to measurement and emphasis on the structural and process dimensions of implementation. Specifically, these tools represent three different applications of two commonly applied observation approaches (i.e., discrete behavioral observation, global rating), while also representing three distinct approaches to measuring the structural and process dimensions of implementation. Table 7 presents an overview of the three direct observation tools, and each specific tool is described in detail next.

OTR. While not a specific tool, discrete behavioral observation approaches have a long history of use with single-case designs, and in studies of specific instructional interactions with students with disabilities (e.g., Smolkowski & Gunn, 2012; Stichter et al., 2006; Stichter et al., 2008; Sutherland et al., 2003). For this study, this approach was

Table 7
Measure Components

Measure	Aspect	Dimension	Approach	Teacher-Student Interaction
OTR	Dosage	Structural Measure	Discrete Behavioral Observation	Opportunities to Respond
CLASS	Quality	Process Measure	Global Rating Scale	Teacher Responsivity Emotional Supports Classroom Organization Instructional Supports
QIDR	Adherence/ Quality/ Participant Responsiveness	Integrated Multicomponent Measure	Global Rating Scale	Quality of Intervention Delivery

Note. OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt.

used to target the structure of one critical discrete teacher-student interaction, *Opportunities to Respond* (OTRs). OTRs are related to student learning in several contexts and across various academic areas (e.g., special and general education; decoding, spelling, math facts, academic engagement) (Smolkowski & Gunn, 2012; Stichter et al., 2006; Stichter et al., 2008; Sutherland et al., 2003; Swanson & O'Connor, 2009). OTRs were defined as any verbal teacher-provided opportunity for students to answer questions, practice content, read aloud, or actively participate in instructional tasks. Based on this definition, OTRs were counted using paper-pencil system for event recording. Coding the frequency of this behavior created an observation that documents the moment-to-moment occurrence of specific instructional interactions (Kennedy, 2005). (See Appendix A for full measure.)

As the OTR observation system is based on a basic observation approach, rather than a specific tool, there are no specific psychometric data available. The teacher-student interactions being observed with this system have been widely examined, however, and have been found to be empirically relevant to student outcomes, particularly the acquisition of basic academic skills (e.g., Smolkowski & Gunn, 2012; Swanson & O'Connor, 2009). The use of discrete behavioral observation, and the use of observation codes based on clearly defined, measurable behaviors have also been shown to reduce rater variability, and increase reliability (Chomat-Mooney et al., 2008; Smolkowski & Gunn, 2012). See the next section for a discussion of training and reliability procedures.

CLASS. The Classroom Assessment Scoring System (CLASS) is a standardized measure of global classroom quality (Pianta et al., 2008). It targets interactional processes found to predict children's academic, social and language development, rather than specific environmental contexts, content or programs, as these other constructs are assumed to impact students only through their interactions with teachers (e.g., Hamre, Goffin, & Kraft-Sayre, 2009, Mashburn et al., 2008). The CLASS has been used to assess prekindergarten through 12th grade classrooms, and early childhood versions are currently under development. It has also been successfully used in classrooms with diverse populations (e.g., ELLs, students with special needs), however it has not been specifically validated for use in classrooms that specifically support these types of learners (e.g., self-contained special education classrooms) (Hamre et al., 2009).

The CLASS is a paper-pencil tool that assesses three broad domains of classroom quality (*Emotional Supports*, *Classroom Organization*, and *Instructional Supports*)

considered to capture the nature of those teacher-student interactions most likely to contribute to student development (Pianta & Hamre, 2009, p. 113). The K-3 version of the CLASS evaluates three to four dimensions per domain. Each dimension is rated on a scale of 1 to 7, with a 1 indicating that the dimension is minimally characteristic, and a 7 indicating that the dimension is highly characteristic of an observed classroom. Low scores (1, 2), mid-range scores (3, 4, 5), and high scores (6,7) are generally described (e.g., a 3 indicates "The middle-range description mostly fits the classroom, but there are one or two indicators in the low range," while a 4 indicates "The middle-range description fits the classroom very well. All, or almost all, relevant indicators in the middle range are present."), and then a more specific example of low, mid range, or high scores is given for each dimension. For instance, on the transition item under the dimension scoring productivity, a low (1, 2) score = "Transitions are too long, too frequent, and/or inefficient," a mid-range (3, 4, 5) score = "Transitions sometimes take too long or are too frequent and inefficient," and a high score (6, 7) = "Transitions are quick and efficient" (Pianta et al., 2008; p. 51). (See Appendix B for overview of dimensions.)

In assessing *Emotional Supports*, the K-3 CLASS measures (a) Classroom Climate (Positive and Negative), (b) Teacher Sensitivity, and (c) Regard for Student Perspectives. Classroom Climate indicates the emotional tone of the classroom and assesses the warmth, respect, enjoyment and enthusiasm displayed in teacher-student connections. Teacher Sensitivity evaluates the teacher's awareness and responsiveness to students' academic and emotional needs. Regard for Student Perspectives assesses the degree to which classrooms are teacher- or student-driven, indicated by teacher flexibility

and willingness to respond to student-initiated activities, and to incorporate students' interests into classroom processes. The *Classroom Organization* domain includes (a) Behavior Management, (b) Productivity, and (c) Instructional Learning Formats. Behavior Management addresses teachers' ability to prevent and redirect misbehavior. Productivity assesses teacher efficiency, or how well instructional time and routines are managed in order to maximize instructional time. The Instructional Learning Formats dimension evaluates teachers' ability to maximize student engagement. The *Instructional Supports* domain measures (a) Concept Development, (b) Quality of Feedback, and (c) Language Modeling. Concept Development considers whether the instructional activities promote higher order thinking or fact-based learning. The Quality of Feedback dimension evaluates whether teacher feedback is focused on formative (e.g., expanding understanding) or summative (e.g., correctness) evaluation. Language Modeling indicates the amount and quality of teachers' language use across classroom interactions with students. (Hamre et al., 2009, p. 15; La Paro et al., 2009).

The CLASS is empirically supported, and has been evaluated using several large national and regional studies. The three broad domains of the CLASS have been examined across a wide range of preschool to fifth grade classrooms, and this three-factor model has typically been recommended and adopted as the best structure for modeling the natural variation in classrooms (Pianta & Hamre, 2009). The constrained sample size in the current study limited the number of scales that could be entered into the model. As such, a general factor (e.g., overall classroom quality/teacher responsivity) that included all ten items was used in the current analysis (Hamre, Pianta, Hatfield, & Jamil, 2014; McGinty, Justice, Piasta, Kaderavek, & Fan, 2012).

While several variations on the CLASS factors have been presented in the literature (e.g., Hamre et al., 2014; La Paro et al., 2011), the traditional three-factor approach has been the most widely accepted, and its psychometric properties are the most widely reported. Hamre and colleagues (2007) present the most comprehensive overview and highlight each factor across year, with the following results across kindergarten settings: Emotional Supports ($\alpha = 0.79$), Classroom Organization ($\alpha = 0.79$), Instructional Support ($\alpha = 0.86$). In terms of reliability, several analyses indicate that trained coders typically reach an average of 87% inter-rater agreement (within 1 score on 7 point scale) with "gold standard" coders (Pianta et al., 2008). The CLASS has been found to have adequate criterion and predictive validity, and is associated with improvements in both academic and social outcomes (Hamre et al., 2007).

QIDR. The Examining Quality of Intervention Delivery and Receipt (QIDR) is a new global rating observation tool that takes an integrated approach to measuring both structural and process-oriented aspects of teacher-student interactions during implementation (Harn, 2013). This tool examines both teacher and student behaviors independently, as well as interactive teaching practices with the recognition that both teachers and students contribute critically to the process of learning. The QIDR has been developed specifically for use with small group interventions, and is designed to be content-independent (e.g., reading, math). It examines both the structure and process of teacher-student interactions related to effective instruction, particularly for students at-risk or with special needs.

The QIDR is a paper-pencil tool that examines three broad constructs (*Quality of Intervention Delivery*, *Overall Intervention Delivery*, *Student Response During Delivery*)

considered to represent those instructional interactions most likely to contribute to the delivery and receipt of effective instruction. Items on the majority of the scales are rated on a Likert-like scale of 0 to three, with a 0 indicating that the item was not implemented, a 1 indicating inconsistent implementation, 2 indicating effective implementation, and a 3 indicating expert implementation. The *Overall Intervention Delivery* scale is a single item that is rated from 0 to 10, with scores of 0-1 indicating that the intervention delivery was ineffective, a 5-6 indicating delivery was proficient, and a 9-10 indicating delivery was highly effective. One item on the *Student Response* scale requires a dichotomous rating indicating whether students were responsive or non-responsive to instruction. To support consistent usage of the QIDR, a detailed rubric has been developed which includes examples and specific scale-level descriptions for each item. For example, for the item rating the quality of teacher modeling, a 0 = "teacher *does not* clearly demonstrate skills/strategies prior to student practice opportunities", a 1 = "teacher *occasionally* clearly demonstrates skills/strategies prior to student practice opportunities", a 2 = "teacher *typically* clearly demonstrates skills/strategies prior to student practice opportunities OR no modeling is used but all students are successful with activities", while a 3 = "teacher *consistently* demonstrates skills/strategies *prior* to student practice opportunities". (See Appendix C for full rubric.)

The *Quality of Intervention Delivery* scale evaluates teacher-driven instructional interactions, and is the longest scale, with 15 items. The first four items address instructional organization and routines: (a) teacher is familiar with the lesson, (b) instructional materials are organized, (c) transitions are efficient and smooth, and (d) teacher expectations are clearly communicated and understood by students. Three items

evaluate emotional and behavioral supports: (e) teacher positively reinforces correct responses and behavior, (f) teacher appropriately responds to problem behaviors, and (g) teacher is responsive to the emotional needs of the students. The next four items are related to specific teaching behaviors that are drawn from the teacher effectiveness literature: (h) teacher uses clear and consistent lesson wording, (i) teacher uses clear signals, (j) teacher models skills and strategies, (k) teacher uses a clear and consistent error correction. Finally, four items evaluate the responsiveness of teacher delivery: (l) teacher provides a range of systematic group opportunities to respond, (m) teacher presents individual turns systematically, (n) teacher modulates lesson pacing, (o) teacher ensures students are firm on content prior to moving forward.

The second scale, *Overall Intervention Delivery*, is a single item that evaluates the overall effectiveness of instruction. This item measures overall effectiveness based on consideration of several factors: quality of delivery, the teacher's understanding of the program, instructional and behavior management, and student engagement and is rated on a scale of 0 to 10. The final scale, *Student Response During Delivery*, is made up of two subscales: Group Student Behavior and Individual Student Response. Group Student Behavior evaluates the overall response of the small group across four items: (a) students are familiar with group routines, (b) students are actively engaged with the lesson, (c) students follow teacher directions, and (d) students are emotionally engaged with the teacher. These items indicate how well the students *as a group* respond to the intervention delivery. The second subscale, Individual Student Response, on the other hand, evaluates specific student responses across three items: (a) emotional engagement, (b) self-regulated behavior, and (c) responsiveness. These items indicate whether specific

students demonstrate particular strengths or weaknesses across these constructs, which are thought to impact student learning.

As with the CLASS, the sample size constraints in the current study limited the number of QIDR scales that could be analyzed. As such, only the combined total of the 15 item integrated *Quality of Intervention Delivery* scale and the four item Group Student Behavior subscale of the *Student Response During Delivery* scale were included in the final analysis. In terms of psychometric properties, the QIDR is a new tool, and in-depth psychometric data are therefore not available. However, initial analyses of inter-rater reliability (IRR) across training observations indicated that IRR was generally good, with average-measure ICC values of .64 across training videos (Cichetti, 1994; Hallgren, 2012; Harn, Spear, Fritz, Berg, & Basaraba, 2014).

Adherence measure. A basic adherence measure was also included as a control, in addition to the three core implementation measures. Given the fact that both schools' implementation of tier two interventions are considered to be fully operational, and that the schools are using slightly different explicit instruction programs, the adherence measure focused on the common approaches to delivering explicit, systematic programs.

This adherence checklist is a paper-pencil tool that evaluates six basic features of explicit instruction: (a) teacher/material preparation, (b) clear and consistent language use, (c) frequent practice opportunities, (d) the delivery of effective corrective feedback, (e) brisk lesson pacing, and (f) the use of clear signaling (Archer & Hughes, 2011). Each component of explicit instruction is rated dichotomously as present or not present greater than 80 percent of the time (See Appendix D for full measure). This measure focuses on basic adherence to explicit instruction program procedures, and was based on tools

commonly used in school settings (e.g., Beil, n.d.; Fritz, n.d.; Heartland Area Education Agency, 2008; *Implementation*, n.d.) that were modified for this study to be program independent. Psychometric data are not available.

Implementation Measurement Procedures

Implementation observation procedures. As previously mentioned, each group videotaped implementation weekly for seven to nine weeks. One interventionist taught two separate groups, leading to a video data set of 64 videos. For the purposes of this study, these videos were coded by three separate cohorts of trained observers using each tool (OTR, CLASS, QIDR). This led to a final data set that contained three distinct measures of implementation across time.

Coding procedures. Each observation tool was coded separately, and had distinct requirements (e.g., classroom experience for the CLASS) so separate coding cohorts were recruited for each tool. Each tool required a group of at least four core coders for reliability purposes (Semmelroth & Johnson, 2013), though more coders were used whenever possible. Undergraduate and graduate students were the main coders, though faculty members of the research team that developed the QIDR were also included as part of that coding cohort. As the principle investigator (PI), I also served as the "master coder" and coded a sample of videos across all three tools to ensure that a measure of stability was provided across the different coding schemes. The OTR, CLASS, and QIDR had cohorts of six, five, and eight coders, respectively.

Training procedures. Each cohort of coders was trained to use their assigned tool by "gold standard" coders (i.e., coders involved in the development of the tool, or who have been trained on the tool by specified experts) in separate training sessions prior to

the start of coding. Each session provided an in-depth introduction to the target observation tool's content, scales, and measurement protocol, and then involved practice sessions with training videos that have been selected from the database, or with similar videos of explicit small group instruction, depending on the amount of training required for each tool. After each training session, coders were required to independently code check-out videos and obtain adequate reliability (two-way random, absolute, single-measure ICC of .6 or higher, Cichetti, 1994; Hallgren, 2012) with a predetermined "true score". Training and overall reliability scores for each tool are presented in Table 8. Specific training procedures for each tool are outlined next.

For the OTR approach, coders were trained in three separate groups due to scheduling and training needs. All coders (n = 6) received one initial two-hour training session that included ample practice time with samples of direct instruction videos. Coders then either proceeded to an independent coding practice round, or were able to attend another two hour training session, depending on their comfort and fluency with coding during the practice opportunities. All coders were required to score two independent practice check-out videos before moving on to independent coding. The true scores for the OTR check-out videos were scored by the PI. All coders obtained adequate reliability after coding two independent practice check-out videos, however one group of two coders had significant variability, and were close to the ICC cutoff (ICC = .62). As such these coders received an additional hour of training where the video intervals with significant variability were discussed and clarified. These coders then coded two additional independent practice check-out videos and obtained strong reliability (ICC =

.89). Overall, five training videos were used across the three training groups, and adequate reliability was obtained by each group (ICCs ranged from .62 to .96).

For the CLASS, coders were trained in one three-hour session by a certified CLASS coder. Coders (n = 5) were introduced to the three broad dimensions of the CLASS, and discussed and viewed video examples of each of the ten items that are scored across the tool. After reviewing the entire tool, coders scored two practice videos over the next week, and scores were compared to the trainer's scores, and areas of disagreement and confusion were discussed and clarified in a one-hour follow-up session. All coders then scored three independent practice check-out videos for the CLASS, given the complexity of the tool, before moving on to independent coding. The true scores for the CLASS check-out videos were scored by the CLASS trainer. Reliability was examined in two ways for the CLASS. Typical CLASS training protocol requires that coders be within plus/minus one score of the true score 80% of the time. This method was used, along with the study-specific reliability protocol to ensure that coders were in line with the typical CLASS framework. All coders obtained adequate reliability on the three independent practice checkout videos (two-way random, absolute, single-measure ICC = .87; overall plus/minus one agreement = 91%, with a range of 83-97%).

For the QIDR, coders were trained in two 3-hour sessions by two core members of the initial QIDR development team. Coders were introduced to each scale of the QIDR individually, and the larger Quality of Intervention Delivery subscale was introduced in subsections; related items were reviewed (e.g., teacher's familiarity with lesson materials, organization of materials) with a chance to view and practice coding samples from videos illustrating these examples. After the training, coders scored 3 independent practice

check-out videos. The true scores for the QIDR videos were developed by the initial QIDR team, who independently coded each training video, discussed any disagreements, and used a consensus-building approach to determine the true score for that video. Reliability was somewhat variable after the first two videos and below the cutoff score (ICC = .51), so coders met with the trainers and discussed agreements and disagreements. Trainees then coded a final practice video and obtained adequate reliability (two-way random, absolute, single-measure ICC = .64).

Video assignment. Each video was randomly assigned an ID number prior to the start of coding to ensure that coders were blind to school, group, and date of delivery. Videos were then randomly assigned to each cohort of coders. Across all three tools, five videos were coded for training purposes, with the remaining set of videos being distributed across each varying cohort depending on size and coder availability (i.e., 9-15 videos per coder for the OTR; 15-16 videos per coder for the CLASS; 10-11 videos per coder for the QIDR). After training, each coder was given a file containing all of their assigned videos for independent coding. Coders were assigned three to four videos per week, and uploaded their scores to a specific Qualtrics survey for each different tool at the end of each week. Coder notes were maintained and collected at the end of the study to ensure that any questions on specific decisions can be examined at a later date.

IRR. Inter-rater reliability was assessed on a randomly selected 30% (n = 19) of the videos for each observation tool. As each tool had a different number of coders, reliability coding was assigned based on each tool. For the OTR and CLASS, there were six and five coders, respectively, so reliability was randomly assigned so that each coder observed an additional three to four videos for reliability purposes. With the QIDR, there

were eight coders, which allowed for a different approach as the number of coders matched the number of videos per group. As such, with the QIDR, coders were systematically assigned to two to three videos for reliability purposes to ensure that no pair of coders shared more than one reliability video assignment, which served as an added measure to remove coder effects from the process. This was not logistically possible with the other tools, due to the number of coders, but was deemed particularly important given the fact the QIDR is a new tool.

IRR was measured across each tool using one-way random, absolute, average-measure ICCs , as not all coders observed each video, and the data for all three measures included interval variables (Gwet, 2012; Hallgren, 2012). ICC values between .6 and .74 are considered good, and .75 and 1.0 are considered to be excellent (Cicchetti, 1994), and as variables that are not perfectly continuous, such as the scales used across the CLASS and QIDR, artificially decrease IRR values, ICCs of .6 and above were accepted.

Reliability checks were held weekly throughout the coding process, to ensure that overall IRR was above .6 each week. Coders were not told which of their video assignments were reliability checks, though the researcher discussed any individual reliability observations that dropped below .7 with individual coders to ensure that coders remained accurate (Kennedy, 2005). Overall IRR for the OTR approach was .94, with a range of .93-.96. For the CLASS, overall reliability was .90, with a range of .86-.95. For the QIDR, overall reliability was .81, with a range of .76-.82. In the final data set, reliability ratings are averaged for each IRR video, so that each video in the data set has one score.

Table 8
Overview of Reliability

Measure	Final Training Reliability: Two-way random, absolute, single-measure ICCs	Overall Reliability: One-way random, absolute, average-measure ICCs
OTR		.94
Group 1	.89	
Group 2	.88	
Group 3	.96	
CLASS	.87	.90
QIDR	.64	.81

Adherence measure. The adherence measure was gathered on a randomly selected 30% of videos by a subset of coders of the other tools (n=5). After a 45-minute training session introducing coders to the adherence checklist, coders were assigned videos following the same procedures outlined above. All videos were double coded for reliability purposes. Given the small sample size and limited range of the adherence measure scale, ICC measures were inappropriate for capturing reliability among coders. A +/-1 agreement method was used instead, and overall IRR was 74% across the subsample of videos.

Confidentiality. Informed consent was collected from all student and instructional assistant participants at the start of the Super K project, however additional measures were taken to protect participant confidentiality throughout this project. Videos were de-identified to ensure that any identifying information (e.g., student names, ELL or IEP status, instructional assistant names) cannot be linked to participants in the student measure or implementation measure data sets. Additionally, all coders completed CITI training and signed an additional confidentiality agreement of not sharing videos, tools,

and data from the project. All coder records will be collected and destroyed once analysis is complete, and videos were deleted from coder computers at the end of the project.

Coding issues. Several videos across the data set had minor audio issues that made coding more difficult, particularly with the OTR tool, in which coding focused on discrete, sentence-level verbal information. Coders were instructed to note and report such issues, and to continue with coding unless less than 50% of the video was audible. If audio issues were reported on a reliability video, individual reliability was checked, and if it dropped below the cutoff (ICC = .6) that video was not used for reliability purposes. Only one coder reported that a video was "uncodable" due to auditory issues. This was a reliability video, so only the primary coder's scores were used, and the secondary coder was assigned to an additional randomly selected video for reliability purposes. ICC levels remained high on all other reliability videos with audio issues.

Experimental Design and Analytic Approach

The current study examined the relationships between three measures of instructional implementation, and their association with an academic outcome measure for at-risk kindergarten students. Given that instructional implementation occurred at the group level, and our interest in examining effects at the student-level and across time, multilevel modeling was used to answer the proposed research questions. Prior to the start of analysis, a cluster randomized trial with cluster-level outcomes power analysis was conducted using the Optimal Design Plus (OD; Raudenbush et al., 2011) software package. This indicated that with the current sample of 31 students (i) nested in 8 groups (j), and α fixed at .05, the likelihood of detecting a statistically significant effect would require an effect size of 2.0 or higher, with .8 power, or an effect size of 1.5 with power

of .42. The underpowered nature of this study is a known limitation. While a well-powered study is clearly more desirable, there is, however, support in the literature for the use of underpowered studies in the initial exploratory phases of research, when potential important relationships are first being explored (e.g., Maxwell, 2004). This study also examines clustered data in a natural school setting, which can be difficult to obtain on a large scale in terms of financial, human, and organizational resources. Research of this type is therefore often underpowered, but can be used to offer important exploratory insights into complex educational processes (Tomcho & Foels, 2012). The current study is therefore considered as an initial, exploratory research project that will be used to provide insight into the relationships between implementation measures and student outcomes, as well as to inform future research about the sample sizes and power needed to truly explore the effects of implementation on student outcomes.

Given these power restrictions, which limited our ability to fit these data to a sophisticated unified model, we took a more stepwise approach and conducted a series of follow-up analyses to examine the full range of research questions (Hox, 2010; Snijders & Bosker, 1999). While this is a known limitation, the nested structure of the data warranted a multilevel approach, and this method may offer more insight than a typical ordinary least squares approach, which tend to inflate Type I errors by underestimating standard errors (e.g., Luke, 2004; Pedhazur, 1997; Raudenbush & Bryk, 2002). These considerations, combined with the exploratory nature of this study, and the important insights this comparison of implementation measures may afford the field make this a reasonable approach. The next section describes the analysis approach for the study.

Model building. The main focus of this study was the relationship between implementation, which occurred at the group-level, and individual student outcomes. Given our interest in level two factors (implementation across groups), the constrained sample size, and the limited parameters available, our models contained only level two predictors (Luke, 2004; Raudenbush & Bryk, 2002). Questions about the relationship between level one student variables and group-level implementation were addressed in a correlational analysis. The first research question, which assessed the degree of association between the three implementation tools, was also addressed using this approach.

The second and third research questions, which examined associations between the measures of implementation and student outcomes, involved both student- and group-level data. To address these questions, we used 2-level hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) with level 1, student (i), nested within level 2, group (j), to account for variance in implementation at the group level, while controlling for differences in student outcomes that can be attributed to group membership. This allowed us to examine the outcome, winter WAT score (Y_{ij}) which was defined as number of correctly decoded nonsense words for student i nested in group j.

The fourth research question assessed implementation over time, and only involved patterns of implementation across time, as student-level data were not collected at multiple time points. As such a 2-level linear growth model was used, where we examined implementation at time (i) across group (j) with each tool. In these models, we were able to examine implementation scores over time by tool (e.g., Y_{ij} = OTR) as the outcome in relation to variability across time for each group. The model building process

is described in detail next. All analyses were conducted using SPSS 21 and 22 for Macintosh and Windows, and HLM 7 for Windows (Scientific Software International, Inc., 2012), and all parameters were estimated using restricted maximum likelihood.

Null model. The model-building process began with a null model, or an unconstrained one-way ANOVA model with no level one or level two predictors. While this model would typically be the initial level one model, based on modeling conventions, sample size constraints limited the amount of predictors we can examine, and as such the null was used as the final level one model. Here, we examined reading achievement when students (level one) were nested in intervention group (level two), and determined the degree to which student reading outcomes depended on group membership (Luke, 2004; Raudenbush & Bryk, 2002). The HLM equations for the final student-level model are described below.

Model 1:

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

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

Mixed Model: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

Means as outcomes. Next, a random intercepts means as outcomes model was constructed, with one level two predictor, implementation effect by tool (W_j). Here, the effect of implementation as measured by each tool (OTR, CLASS, QIDR) was entered into the model as a level two predictor to examine reading achievement when students (level one) were nested in intervention group (level two) while controlling for the level of implementation in each group, by measure. Level two predictors were calculated by taking the average of the repeated measures for each tool (e.g., 8 observations of Group 2

using the CLASS), and entering each as a grand-mean centered term into separate models. Here again our sample size restricted the number of predictors that could be entered into the model, so each tool was modeled separately. This also limited the number of random parameters that could be entered into the model, and as there were no level one predictors, only the intercepts were allowed to vary randomly. The equations for these parallel group-level models are presented below.

Models 2-4:

$$\text{Mixed Model: } Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + r_{ij}$$

$$\text{Model 2: } Y_{ij} = \gamma_{00} + \gamma_{01} OTRAvg_j + u_{0j} + r_{ij}$$

$$\text{Model 3: } Y_{ij} = \gamma_{00} + \gamma_{01} CLASSAvg_j + u_{0j} + r_{ij}$$

$$\text{Model 4: } Y_{ij} = \gamma_{00} + \gamma_{01} QIDRAvg_j + u_{0j} + r_{ij}$$

Incremental variance explained. After the basic means as outcomes models were examined, a set of extended means as outcomes models were constructed in order to examine the additional variance of each tool helped to explain when added to a model with each other predictor. Here again, the sample size restricted the number of predictors that could be entered into the model, as the size of the highest level determines the number of parameters available for any model (e.g., Snijders, 2005; Snijders & Bosker, 1999). As our highest level, group, had an *n* of 8, a model with all three tools resulted in an overfit model that was uninterpretable. As such, only two implementation tools were compared at a time. As such, to specifically compare the effects of each implementation tool, we examined between-group variance for each pair of tools (OTR and CLASS; CLASS and QIDR; QIDR and OTR). Each model is presented below.

Models 5-7:

Mixed Model: $Y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + u_{0j} + r_{ij}$

Model 5: $Y_{ij} = \gamma_{00} + \gamma_{01}OTRAvg_j + \gamma_{02}CLASSAvg_j + u_{0j} + r_{ij}$

Model 6: $Y_{ij} = \gamma_{00} + \gamma_{01}OTRAvg_j + \gamma_{02}QIDRAvg_j + u_{0j} + r_{ij}$

Model 7: $Y_{ij} = \gamma_{00} + \gamma_{01}CLASSAvg_j + \gamma_{02}QIDRAvg_j + u_{0j} + r_{ij}$

Next, the single predictor and dual predictor models' pseudo-R^2 were examined in order to determine the unique variance of each tool in the dual model added. In other words, the level two pseudo-R^2 was calculated for each model, and compared to the pseudo-R^2 for the original single-measure models (i.e., 2-4) to determine both the unique variance of the pair of tools entered in each model. For example, the level two pseudo-R^2 for models 2 and 3 were subtracted from the level two pseudo-R^2 for model 5 to determine the unique variance of the OTR and CLASS, respectively, in model 5. This process was repeated for models 5-7.

The unique variance of these tools was then examined through a comparison of empirical Bayes (EB) residuals to determine whether specific models did a better or worse job of predicting the eight groups. As a residual, the EB estimate represents the difference between the outcome value that is predicted based on ordinary least squares (OLS) estimation, and a Bayesian estimation that accounts for the number of students in each group. With the OLS estimate, all groups are weighted equally, regardless of size, whereas the EB estimates "shrink" toward the grand mean, where larger groups receive less weight, and smaller groups receive a greater weight. While this "shrinkage" introduces bias into the estimate, it is also considered to be more precise, and is therefore useful in interpreting group differences (Hox, 2010; Raudenbush & Bryk, 2002).

An EB residual for each group was estimated using the initial single-tool models (i.e., Models 2-4). Given the small sample size and limited power, correlations between the EB and OLS estimates for each tool were examined as an additional representation of the relationship between one tool (e.g., QIDR) and the unexplained variance of the other two tools (e.g., the EB_OTR). In addition, bivariate graphs comparing each pair of EB residuals and each residual to the obtained OLS estimates for each tool were examined in a scatterplot matrix in order to examine how the different small groups were ranked according to each tool and the different estimates.

Growth model. Finally, a linear growth model was constructed. This involved a second data set, where scores from all observations across the entire seven-nine weeks of the intervention (n = 64) were included for each tool. As with the earlier models, the small level two sample size restricted the number of parameters available for analysis, so only the linear trends of each outcome were modeled across time. This also restricted any ability to examine the functional form (e.g., quadratic, cubic) of these data, however visual analysis of each group's implementation patterns across time were examined, and a linear relationship was determined to be reasonable. Given these restrictions, only an unconditional growth model was built for each tool. An uncentered time variable, where observation time was coded from zero to eight, was entered into level one of the model, and allowed to vary randomly for each observation tool. This variable was not centered so that the intercept represented the average group score at the first observation across each tool. These final models are presented below.

Models 8-10:

Mixed Model: $Y_{tj} = \beta_{00} + \beta_{01}W_j + r_{0j} + r_{tj} + e_{ti}$

Model 8: $OTR = \beta_{00} + \beta_{01}TIME_j + r_{0j} + r_{1j}TIME_j + e_{ti}$

Model 9: $CLASS = \beta_{00} + \beta_{01}TIME_j + r_{0j} + r_{1j}TIME_j + e_{ti}$

Model 10: $QIDR = \beta_{00} + \beta_{01}TIME_j + r_{0j} + r_{1j}TIME_j + e_{ti}$

CHAPTER IV

RESULTS

Descriptive Analysis

Raw student-level and group level observation data were examined, cleaned, and screened using SPSS 22.0 for Mac prior to running any statistical analyses.

Missingness analysis. Four of the original 35 students in the intervention did not have complete data due to attrition or absences during testing. A missingness analysis was conducted to determine whether any significant differences existed between the original data set of 35 students and the analytic data set of 31 students. Missing data were predicted using the expectation maximization (EM) method, and then analyzed using Little's MCAR test, $\chi^2(7) = 9.09$, $p = .246$. These results indicate that data can be assumed to missing completely at random, and that the analytic data set provides an adequate representation of the sample population.

Descriptives. Tables 9 and 10 provide overviews of descriptive data for student and group-level outcomes. In terms of student-level data, students' average winter standardized WAT score was 99.9 (*SD* = 7.47), with a range of 94-114, which was indicative of the restricted sample of at-risk kindergartners. Implementation data were also examined and revealed several different patterns across each tool. OTR data revealed that teachers provided high numbers of individual and group opportunities to respond across the groups (*M* = 252.56, *SD* = 74.27), though the range of 116-458 indicated there was quite a bit of variability. Overall CLASS scores (M = 3.56, SD = 0.6) were fairly low, with a range of 2.25-5, indicating that no teachers had overall scores in the high (6, 7) range. QIDR Composite scores (*M* = 35.77, *SD* = 10.86), with a range of 14-56,

indicated that as a group, teachers scored slightly above the average QIDR rating. The subsample of videos that were assessed using the Basic Adherence measure (M = 5.63, SD = 0.68), with a range of 4-6, indicated that teachers were strongly adhering to basic program procedures.

Table 9
Child Outcome Descriptives

Child Outcomes (post-test)	*N*	*M*	*SD*	Min	Max
Woodcock Reading Mastery Test - Word Attack SS	31	99.9	7.47	94	114

Note. SS = Standard Score. The data have been cleaned. Missing student cases were excluded if they were missing data for child outcome measures.

Histograms, skewness, and kurtosis were examined for all variables at both levels. WAT scores were positively skewed, with over half of the students in the sample (n = 18) scoring a 94 (raw score = 0) on the winter WAT measure. Floor effects are often common in measures that target new skills, and as such this may be part of the nature of working with a kindergarten sample (e.g., Catts, Petscher, Scnatschneider, Bridge, & Mendoza, 2009), however these measures were taken in the winter of students' kindergarten year, after targeted early literacy instruction. As such, this finding may be more indicative of the at-risk nature of the sample population included in this study. Given these floor effects, a bivariate correlational analysis was run comparing the full sample (n = 31) to a sample where students who received a score of 94 were removed (n = 13). As can be seen in Table 11, correlations between the implementation measures and student outcome scores went up substantially. While these floor effects highlight issues around the normality of the sample distribution, the results of this correlational analysis indicate that the analysis is most likely offering a more conservative estimate of the relationships

Table 10
Implementation Measure Descriptives

Overall		*N*	*M*	*SD*	Min	Max	*ICCs*
OTR		64	252.56	74.27	116.00	458.00	0.94
CLASS		64	3.56	0.60	2.25	5.00	0.90
QIDR		64	35.77	10.86	14.00	56.00	0.81
Adherence		19	5.63	0.68	4.00	6.00	
By Group		*N*	*M*	*SD*	Min	Max	
1	OTR	4	290.71	75.14	162.00	392.50	
	CLASS	4	3.85	0.27	3.50	4.30	
	QIDR	4	46.64	5.39	38.00	53.00	
2	OTR	3	233.50	55.61	163.00	316.00	
	CLASS	3	3.98	0.42	3.60	4.80	
	QIDR	3	42.31	4.61	36.50	49.00	
3	OTR	3	259.94	82.05	167.00	396.50	
	CLASS	3	3.49	0.59	2.80	4.70	
	QIDR	3	41.06	5.09	35.00	51.00	
4	OTR	3	292.31	43.03	238.00	358.00	
	CLASS	3	3.81	0.19	3.50	4.05	
	QIDR	3	37.25	6.87	26.00	50.00	
5	OTR	3	303.81	73.73	216.00	458.00	
	CLASS	3	4.23	0.40	3.70	5.00	
	QIDR	3	47.25	6.18	39.00	56.00	
6	OTR	5	194.00	64.70	116.00	316.00	
	CLASS	5	2.69	0.34	2.25	3.30	
	QIDR	5	22.81	5.79	14.00	28.50	
7	OTR	5	217.13	68.94	150.00	332.00	
	CLASS	5	3.29	0.37	2.85	3.75	
	QIDR	5	26.88	3.14	24.00	32.00	
8	OTR	5	232.88	72.40	144.00	367.00	
	CLASS	5	3.18	0.33	2.80	3.70	
	QIDR	5	22.63	4.47	17.00	28.00	

Note. OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt. ICCs indicate level of observer agreement.

between implementation and this student outcome. This, coupled with the fact that HLM is generally fairly robust to moderate violations of normality (Maas & Hox, 2004), led to the decision to continue with the analysis as planned with the full sample of students included.

Table 11
Bivariate Correlational Analysis of Group Level Differences Between Full and Restricted Sample to Gauge Impact of Floor Effects

	Full Sample WAT Score	Restricted Sample WAT Score
OTRAvg_mean	.289	.793*
CLASSAvg_mean	.262	.687
QIDRAvg_mean	.544	.678

Note. WAT_SS = Word Attack Standard Score; OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt.
Full Sample $n = 31$; Restricted Sample $n = 13$; $*p < .05$

All other data fell within the normal range, with no severe outliers or skew. Examination of bivariate scatterplots comparing level one WAT scores to each implementation measure revealed that there were no significant outliers, and clear differences between groups. Bivariate scatterplots comparing all level two implementation measures indicated that there are clear linear relationships between all three direct observation tools.

Testing of model assumptions. Model adequacy was assessed through an examination of the residuals for each final model (i.e., model four for the single-tool models, model seven for the dual predictor models, and each individual model for the parallel growth models). Despite the strong floor effects on the outcome variable in the models that include student outcomes (i.e., models four and seven), the residuals were independent, and normally distributed. The growth model residuals and predictors were also found to be independent and normally distributed, and the overall patterns across all models indicate that HLM assumptions appear tenable.

Descriptive analysis relating student and implementation variables. Given the sample size restrictions, and the fact that the model could not support the inclusion of any student-level predictors, an initial correlational analysis was conducted to examine the

relationship of important student-level variables (i.e., demographics, pretest data) with group-level implementation, and student outcomes. Bivariate correlations among group-level variables (i.e., student data is aggregated across group) are reported to adjust for inflated scores due to repeated group values. None of the student-level demographic variables or pretest scores were significantly associated with posttest scores. This lack of significance is not surprising given the small number of groups (n = 8), however the patterns of correlation between student demographics and group-level implementation are worth noting. Specifically, ELL status was associated with the OTR (r = .50, p = .21), CLASS (r = .46, p = .26), and QIDR (r = .59, p = .12), while special education status was associated with the OTR (r = .56, p = .15), CLASS (r = .67, p = .07), and QIDR (r = .42, p = .30), respectively. Group pretest scores were not significantly associated with posttest scores or implementation measures. Table 12 summarizes the bivariate correlations between these student- and group-level variables.

Results

Research Question 1: How do three implementation/observation tools relate to each other? Bivariate correlations between the three implementation tools were examined to determine the interrelations between these three measures. All three implementation measures were significantly associated. Specifically, the OTR was related strongly to both the CLASS (r = .82, p < .05) and QIDR (r = .80, p <.05), while the CLASS and QIDR were highly related (r = .90, p <.01). This high degree of multicollinearity among the three implementation measures also suggests that serious estimation issues may arise, particularly in the models with multiple predictors, such that models fail to converge, or produce suppression effects.

The relationship between these three tools and the measure of basic adherence was also examined using bivariate correlations. As the basic adherence measure was collected on a randomly selected 30% of videos, the median value (*Mdn* = 6.00) of these observations was applied to the entire data set, though there was very little variability across these observation scores. Basic adherence was moderately, but insignificantly related to three implementation tools. Table 12 provides an overview of all correlational analyses.

Table 12
Bivariate Group-Level Correlations Between Student-Level Variables and Implementation by Group

	1	2	3	4	5	6	7
Student-Level Variables							
1. Group WAT_SS Posttest	-						
2. Group WAT_SS Pretest	.215	-					
3. Group IEP Status	-.333	.138	-				
4. Group ELL Status	.252	-.003	-.125	-			
Implementation Variables							
5. OTR	.289	-.198	.564	.495	-		
6. CLASS	.262	.337	.670	.457	.822*	-	
7. QIDR	.544	.253	.419	.589	.802*	.894**	-
8. Adherence	.306	.143	.046	.493	.364	.224	.352

Note. WAT_SS = Word Attack Standard Score; IEP = Individualized Education Plan; ELL = English Language Learner; OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt.
$n = 8$; *$p < .05$; **$p < .01$

Research Question 2: How do the observational tools relate to student outcomes? Which observational tool individually accounts for the most variance in student outcomes? Table 13 provides an overview of the results of analyses that examined the impact of the three individual implementation measures in predicting student outcomes (See page108-109 for an overview of model building process). The first, or null, model was used to estimate the amount of variance at each level, with no

predictors in the model. This indicated that there was significant variance at both the student level $t(7) = 48.05$, $p < .001$, and the group level, $\chi^2 = 27.76$, $p < .001$, with 56% of the variance between kindergarten students' reading outcomes occurring at the student level, and 44% of the variance between groups. These results indicated that hierarchical analyses were appropriate, given the large amount of variance that occurred at the group level. In the second model, each group's average OTR score was entered into the model individually. While the intercept and its accompanying variance component remained statistically significant, the coefficient estimating the predictive power of the OTR implementation measure was not, $t(6) = 0.85$, $p = 0.43$. Given the underpowered nature of this study, statistical insignificance is not surprising. Other parameters in each model were therefore examined to explore how well each implementation measure predicted student outcomes. Specifically, for each model the level one and level two ICCs were calculated to examine the amount of variance per level, along with the level two pseudo-R^2, which indicates the amount of level two variance that is explained by the predictors, as well as a comparison of deviance scores as a way to estimate model fit. For the OTR model, the amount of variance at each level was comparable to the null, and while the amount of variance at level two was significant, $\chi^2 = 22.90$, $p < .001$, the amount of level two variance explained by the OTR measure was negligible, pseudo-$R^2 = -.03$, and the deviance score increased from the null, indicating that OTRs, as measured here and with this sample, were a poor predictor of group differences in students' WAT scores.

In the third model, each group's average CLASS score was entered into the model individually. Here again the intercept and its accompanying variance estimate remained significant, however the coefficient for the CLASS was also not significant, $t(6) = 0.76$, p

= .48. Here again the ICCs indicated comparable variance at each level to the null, though the deviance score decreased. These model statistics, combined with the pseudo-R^2 = -.06, indicated that the CLASS is a poor unique predictor of group differences in at-risk kindergarten student's WAT scores.

In the final individual tool model, each group's average QIDR score was entered into the model. Overall patterns of significance remained the same across all three individual models, with the intercept and its random effect still significant, while the coefficient for the QIDR was not significant, $t(6) = 1.78$, $p = 0.13$. Examination of the model statistics, however, indicated a different pattern when the QIDR was the individual predictor. Here, the amount of variance between levels shifted, with 67% of the variance now at level one, and 33% at level two. This level two variance was still significant, $\chi^2 = 16.31$, $p < .01$, but the amount of variance explained by adding the QIDR as a level two predictor was quite different, pseudo-R^2 = .36, indicating that 36% of the variance at level two was accounted for by the QIDR. This, coupled with a decrease in model deviance, indicated that the QIDR, despite its insignificant coefficient, may be a strong predictor of group differences in at-risk kindergarten students' reading outcomes. As such, model four was considered the final individual tool model.

While the results described above indicated that the QIDR was the best individual predictor of group differences in student outcomes, recent research and theory indicate that individual opportunities to respond (INDOTRS) may be a stronger predictor than overall or group OTRs (e.g., Doabler, Baker, Kosty, Clarke, Miller, & Fien, in press; Smolkowski & Gunn, 2012). Given the fact that group opportunities are built into the programs used in this study, and that OTRs were liberally defined in the current study, a

follow-up analysis was conducted where the average number of INDOTRs delivered per group were examined uniquely. While INDOTRs are also built into these intervention programs, it is far more under the control of the teacher as to how many to provide within a given lesson. Each group's average INDOTRs were entered into the model, and results were almost identical to the overall OTR model, t (6) = 0.03, p = 0.65. Model statistics were also comparable, with 53% of the variance in student WAT scores occurring at the student level, and 47% at the group level, though the level two variance explained by including only INDOTR as a unique predictor was again negative, pseudo-R^2 = -.14. Given these findings, average INDOTR were therefore dropped as a predictor and not examined in relation to other tools in favor of developing the most parsimonious model building process.

Research Question 3: How do the observational tools uniquely account for variance in student outcomes when entered into the model simultaneously? Table 13 (see page 123) provides an overview of analyses that examined the incremental variance explained by each set of implementation measures (see page 110 for an overview of model building process). The first dual predictor model involved entering the average OTR and CLASS scores into the model simultaneously. The intercept and its variance component were both significant, but neither the OTR nor CLASS coefficient was a significant predictor of WAT scores, $t(5) = 0.55$, $p = .61$, and $t(5) = 0.18$, $p = .86$, respectively. The amount of variance at level two was significant, $\chi^2 = 22.89$, $p < .001$, with the 50% of the variance between kindergarten students' reading outcomes occurring at the student level, and 50% of the variance between groups, and there was a decrease in deviance from the null model. However, the variance explained at level two, pseudo-R^2 =

-.31, and the fact that neither of these tools were effective individual predictors of group-level differences in student WAT scores, indicated that the increased level two variance was most likely due to a misspecified model with poor predictors (Raudenbush & Bryk, 2002).

In the sixth model, the average OTR and QIDR scores were entered simultaneously. Here again the intercept and its variance component were significant, while the predictor coefficients were not, $t(5) = -0.67$, $p = .53$ for the OTR, and $t(5) = 1.52$, $p = .19$ for the QIDR. The model statistics for this dual model indicated that 63% of the variance in the model was now occurring at the student level, and 37% at the group level. The OTR and QIDR together accounted for 24% of that level two variance (pseudo-$R^2 = .24$), however the deviance increased from the null model, and a comparison this model to the model where the QIDR was entered individually indicated that the OTR was actually decreasing the predictive power of the QIDR.

In the final dual predictor model, average CLASS and QIDR scores were entered into the model simultaneously. Here again only the intercept and its variance component were significant, with nonsignificant outcomes for the CLASS, $t(5) = -1.62$, $p = .17$ and QIDR $t(5) = 2.34$, $p = .07$, though the fact that the coefficient for the QIDR was approaching statistical significance with such an underpowered study is notable. In this model, 73% of the variance occurred at the student level, and only 26% occurred at the group level, but this amount was still significant, $\chi^2 = 11.48$, $p < .05$, and the amount of level two variance accounted for by the QIDR and CLASS together was notably higher, pseudo-$R^2 = .55$, and the deviance of the model also dropped to its lowest point, indicating that model 7 was the best fit for these data and predictors.

Table 13
Fixed and Random Effects Estimates Models WAT Posttest Scores

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Fixed Effects							
Intercept	100.37***	100.44***	100.45***	100.52***	100.45***	100.52***	100.49***
	(2.09)	(2.11)	(2.14)	(1.79)	(2.32)	(1.90)	(1.61)
OTR		0.05			0.04	-0.06	
		(0.85)			(0.11)	(0.09)	
CLASS			3.44		0.93		-12.36
			(4.54)		(8.80)		(7.65)
QIDR				0.33		0.51	0.86
				(0.18)		(0.33)	(0.37)
Random Effects							
Group (intercept)	25.98***	26.66***	27.66***	16.65**	33.93***	19.84**	11.75*
Student residual	33.15	33.34	33.28	33.62	33.30	33.46	33.34
Model Statistics							
ICC – Level 1	0.5606	0.5557	0.5461	0.6688	0.4953	0.6278	0.7394
ICC – Level 2	0.4394	0.4443	0.4539	0.3312	0.5047	0.3722	0.2606
Pseudo R^2							
Level 2		-0.0262	-0.0647	0.3591	-0.3060	0.2363	0.55
Level 1		-0.0057	-0.0039	-0.0142	-0.0045	-0.0094	-0.0057
Deviance	203.23	204.59	195.96	200.14	200.32	204.61	193.59
Parameters	2	2	2	2	2	2	2
Deviance Change	-	1.36	-7.27	-3.09	-2.91	1.38	-11.02

Note. Parentheses denote standard errors. Level 2 predictors are group centered. *p < .05. **p < .01. ***p < .001.

After these models were specified, the pseudo-R^2 were calculated and examined in order to determine the unique variance contributed by each tool in each dual predictor model. Given the results and fit of the OTR/CLASS and OTR/QIDR dual models, these models were dropped in favor of a more parsimonious analysis, as they were not explaining more meaningful variance than the QIDR alone. In the CLASS/QIDR model, which accounted for 55% of the variance at level two (26%), the CLASS was found to account for 19% of the variance uniquely, while the QIDR accounted for 36%.

It is worth noting that this examination of the unique variance that the CLASS accounted for in addition to the QIDR could not be reversed (i.e., a full commonality analysis examining shared and unique variance attributed by both tools) due to the fact that the CLASS increased level two variance when entered into the model individually (see Model 3). This led to a negative Pseudo-R^2, which indicates model instability. This, coupled with the fact that the sign for the CLASS coefficient changed from positive to negative in Model 7, suggests that suppression is in operation. These findings will be discussed in depth in the next chapter, however overall this suggests that Model 7 may be oversaturated, or overfit, with too many predictors (Raudenbush & Bryk, 2002).

The unique variance of these tools was further examined through a series of comparisons of the EB residuals. Here, all tools were included, as this comparison was conducted in part to determine whether there were meaningful relationships between the tools that were not being detected due to the limited power in the analysis. As was expected given the high bivariate correlations between the three tools, their EB residuals were also significantly correlated. More interestingly, the QIDR tool was found to be moderately positively, but not significantly, related to both the OTR and CLASS EB

residuals, $r = .33$, $p = .42$, and $r = .33$, $p = .43$, respectively, which may indicate that both tools, when used with the QIDR, might account for more variance in a well-powered analysis. Table 14 presents an overview of these findings.

The final component of the analysis for the third research question involved visually examining a matrix scatterplot that compared all the EB residuals and the average tool scores. This visual analysis allowed for an examination of individual groups, and how the different measures and estimates impact predicted group implementation. This analysis indicated that overall patterns of estimation were consistent across the groups, however one group, group five, was consistently overestimated in the OLS (i.e., average tool) models. This group had only three students, and further inspection indicated that all students scored at the floor, while the teacher here was rated highly across all three implementation measures. As such, the EB residual models indicated that this was leading to an inflated estimate for this group. All other group patterns appear to be fairly consistent across tools. Figure 2 provides an overview of these findings.

Research Question 4: What does implementation look like across intervention time by implementation tool? Table 15 presents an overview of the growth analyses examining implementation across time (See page 112 for an overview of model building process). Each tool was modeled separately in an unconditional growth model, providing basic information around patterns of growth and whether implementation varied significantly between groups across time. In the first model, each group's OTR scores were the outcome, and time was entered as an uncentered predictor at level one, and allowed to vary randomly at level two. The intercept was significant, $t(7) = 16.89$, $p < .001$, however its random component was not, $\chi^2 = 10.57$, $p = .16$.

Table 14

Bivariate Correlations between Empirical Bayes Residuals and Average Group Score by Tool

	1	2	3	4	5	6
Empirical Bayes Residuals						
1. EB_OTR	-					
2. EB_CLASS	.98**	-				
3. EB_QIDR	.92**	.94**	-			
Implementation Variables						
4. OTR	.00	.09	-.17	-		
5. CLASS	.03	.00	-.27	.82*	-	
6. QIDR	.33	.33	.00	.802*	.89**	-

Note. EB = Empirical Bayes; OTR = Opportunities to Respond; CLASS = Classroom Assessment Scoring System; QIDR = Examining Quality of Intervention Delivery and Receipt. n = 8; *p < .05; **p <.01

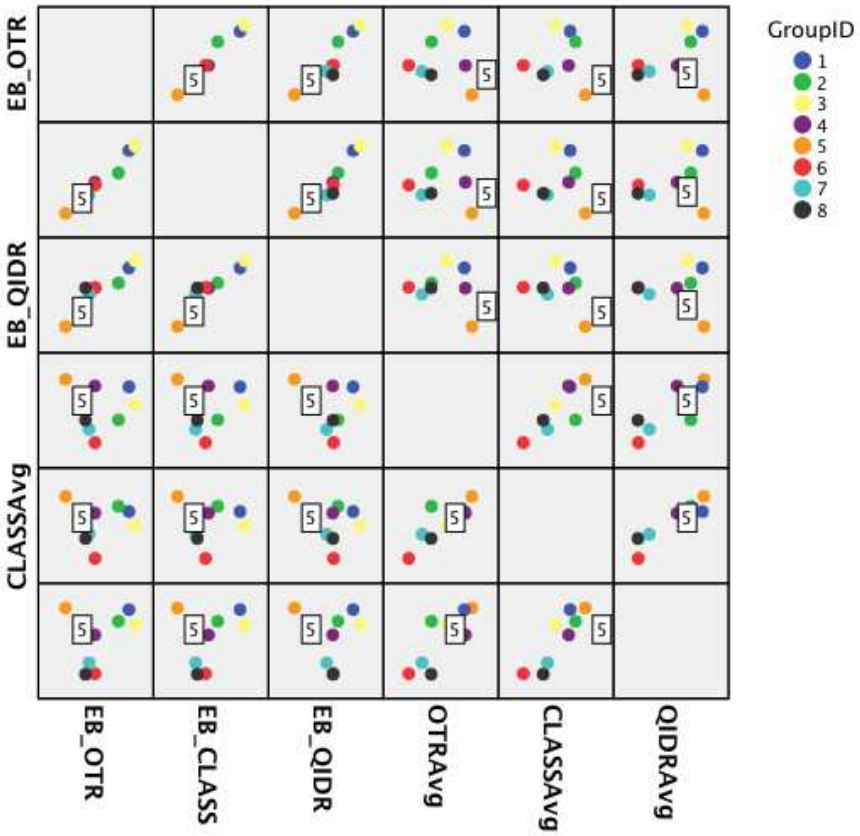

Figure 2. Matrix scatterplot of empirical Bayes (EB) residuals by average tool scores. Ticks on the Y-axis represent the average QIDR scores, average CLASS scores, average OTR scores, and then the QIDR, CLASS, and OTR EB residuals, from bottom to top.

Conversely, the fixed effect for time was not significant, $t(7) = -0.62$, $p = .56$, however its random effect was significant, $\chi^2 = 29.07$, $p < .001$. These results indicate that while groups' average OTR scores are different from zero at observation time one (i.e., the intercept), there was no significant growth overall, though there were significant differences in growth between individual groups across time. In other words, there was no significant growth on average, but individual groups varied in their growth patterns across time. An examination of the model statistics indicated that 22% of the variance in the model was between groups, and that 27% of that variance was attributable to differences in trends over time between the groups. Figure 3 represents each group's growth patterns using the OTR measure.

The second model used the same process to examine implementation as measured by the CLASS across time. Here, the intercept and its random component were both significant, $t(7) = 25.32$, $p < .001$, and $\chi^2 = 19.52$, $p < .01$, respectively, while the fixed effect for time was not significant. The random effect for time closely approached significance, $\chi^2 = 13.21$, $p = .07$, indicating that here again individual groups' growth patterns across time were notable, though overall growth was not significant. The model statistics here indicated that 44% of the variance in the CLASS time model was between groups, but that only 3% of that variance was attributable to differences in change over time between the groups. Figure 4 represents each group's growth patterns using the CLASS.

The QIDR was examined in the final growth model. The intercept and its variance component were both significant, $t(7) = 10.97$, $p < .001$, and $\chi^2 = 51.38$, $p < .001$,

Table 15
Unconditional Growth Models Examining Implementation Across Time by Tool

Parameter	Model 8	Model 9	Model 10
Fixed Effects			
Intercept	268.85***	3.56***	36.40***
	(-15.92)	(0.14)	(3.06)
Time	-3.95	0.002	-0.15
	(6.41)	(0.03)	(0.32)
Random Effects			
Residual, e_{ti}	3248.68	0.13	27.60
Group (intercept), r_{0i}	673.23	0.10***	76.59***
Group (time), r_{1i}	250.43***	0.003	0.16
Model Statistics			
ICC, (Group)	0.2214	0.4421	0.7355
Deviance	715.10	81.80	415.68
Parameters	4	4	4

Note. Parentheses denote standard errors.
*p<.05. ** p<.01. ***p<.001

respectively, however neither the fixed, $t(7) = -0.47$, $p = .65$ nor random components, $\chi^2 = 6.90$, $p > .50$, for time were significant. The QIDR time model accounted for 74% of the variance between groups, however less than 1% of that variance was attributable to differences in trends over time between groups. Figure 5 represents each group's growth patterns using the QIDR. As can be seen here, while groups started at significantly different points, there was no significant growth or downward trends, despite notable variability within groups.

Finally, given the theoretical and empirical support for the predictive power of individual opportunities to respond (e.g., Doabler et al., in press; Smolkowski & Gunn, 2012), the number of INDOTRs delivered to each group across time was again examined in a follow-up analysis to determine whether there was a differential effect for individual or overall OTRs. The results of this fourth model were highly similar to the overall OTR

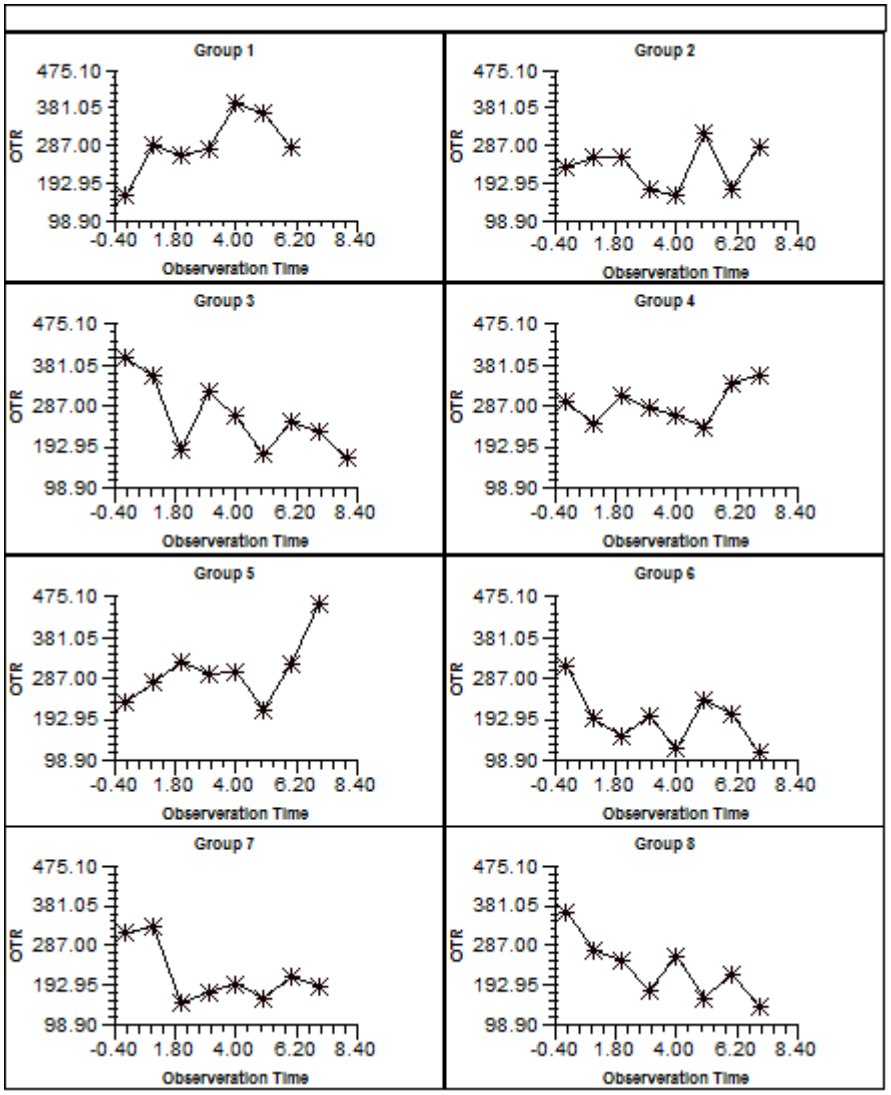

Figure 3. OTR measure growth patterns across observation time by group.

model, where the intercept was significant, $t(7) = 11.41$, $p < .001$ and the fixed effect for time was not significant, $t(7) = -0.83$, $p = .43$, however the intercept's random component approached significance $\chi^2 = 13.72$, $p = .06$, and the random effect for time was again significant, $\chi^2 = 42.39$, $p < .001$. As with the overall OTR measure, there was no significant growth overall, however individual groups varied significantly, and here each group's individual OTRs were almost significantly different from the average at the first

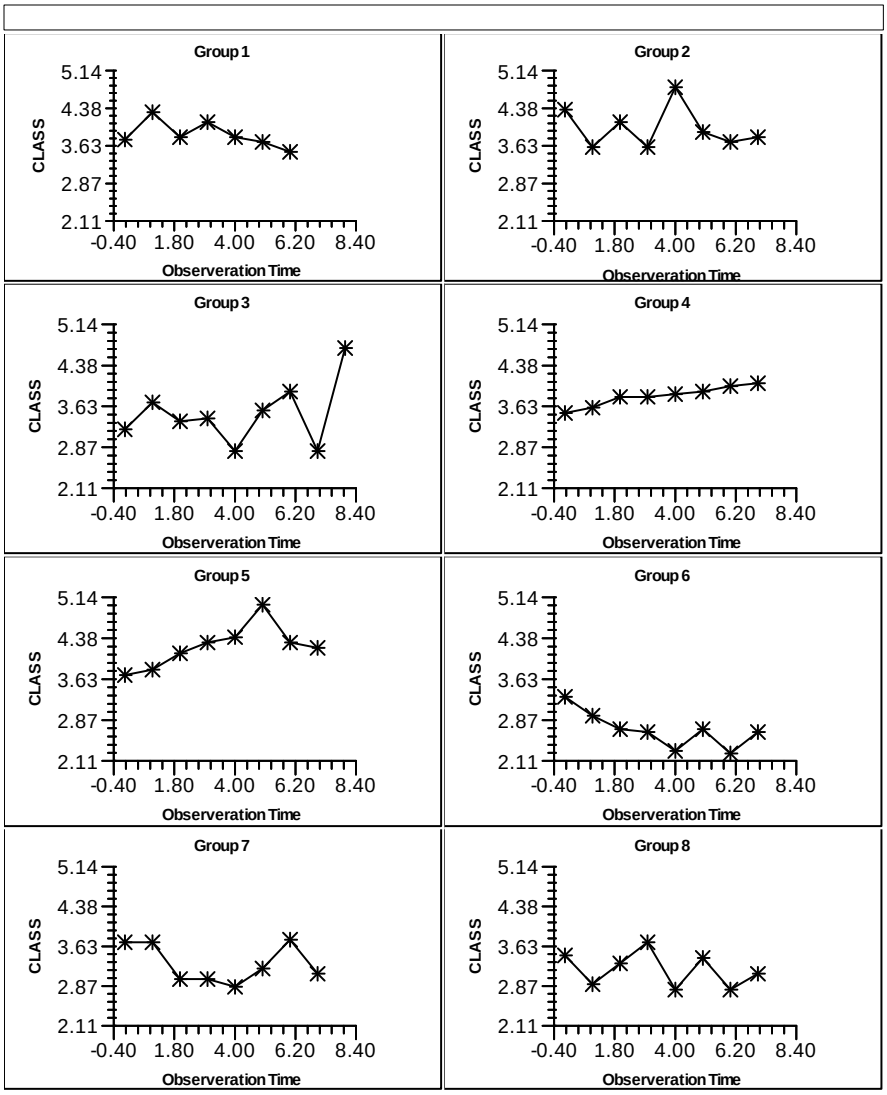

Figure 4. CLASS growth patterns across observation time by group.

observation point (i.e., the intercept). Figure 6 represents each group's growth patterns using only INDOTRs.

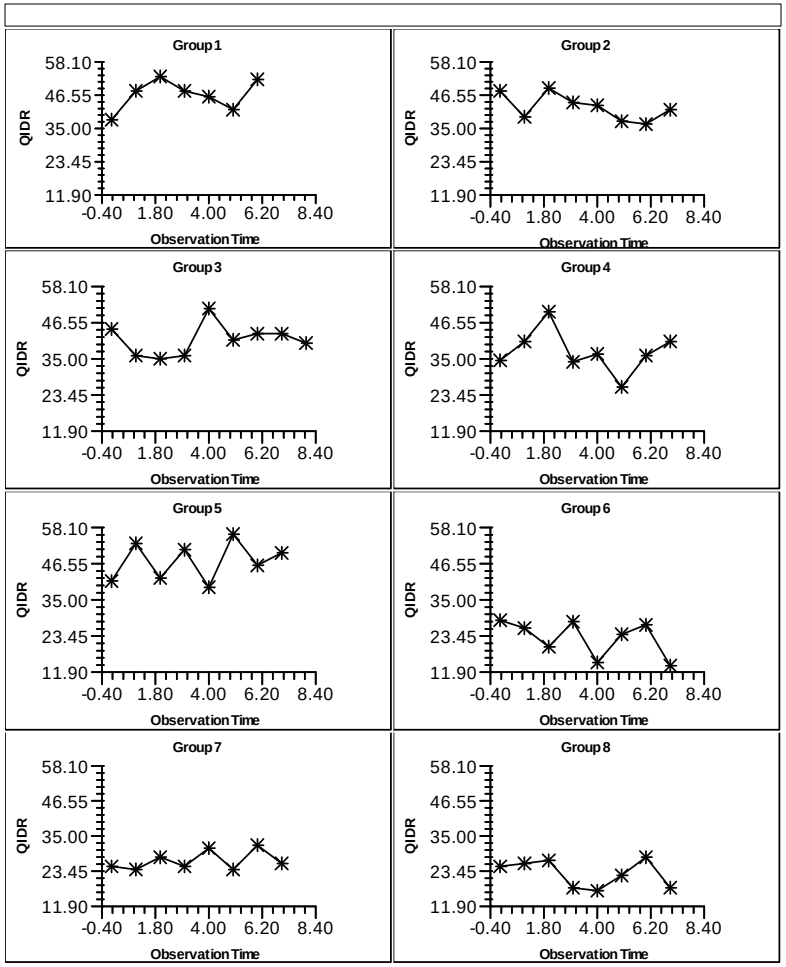

Figure 5. QIDR growth patterns across observation time by group.

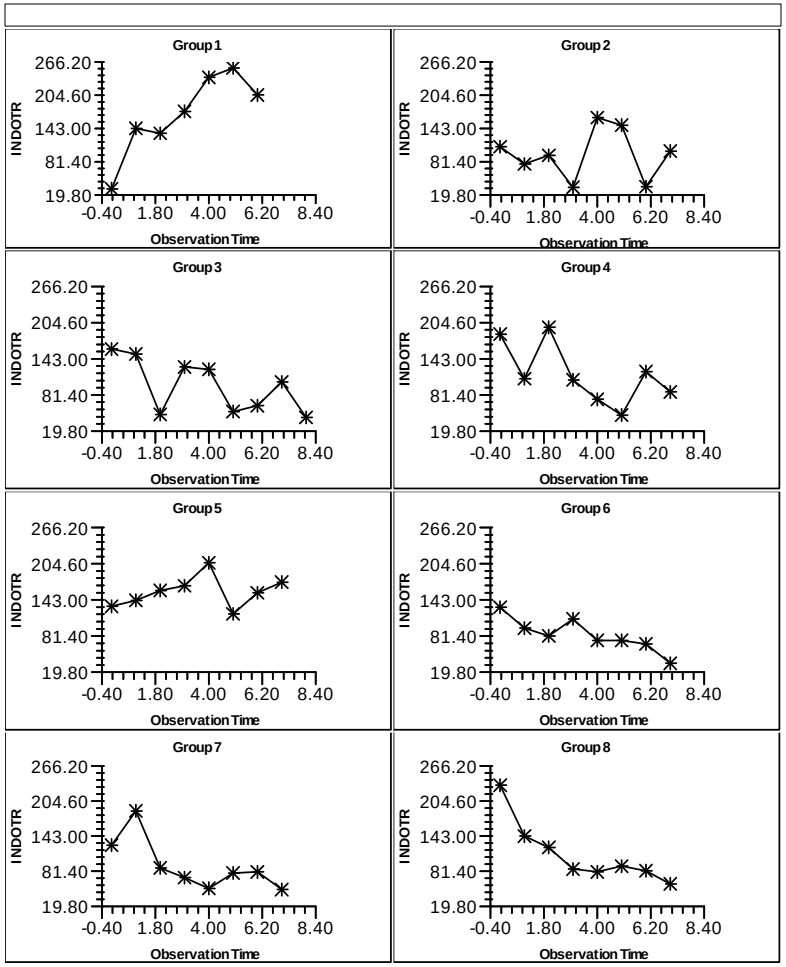

Figure 6. Individual OTR measure growth patterns across observation time by group.

CHAPTER V

DISCUSSION

The purpose of the current study was to compare three measures of instructional implementation to determine how these tools relate to each other and to student outcomes for at-risk kindergarten students receiving small group reading interventions. Patterns of implementation across time, as measured by each tool, were also examined. Overall, results of this analysis indicated that (a) the implementation tools were highly correlated with each other, (b) only the QIDR independently accounted for group differences in WAT scores, (c) together the QIDR and the CLASS appear to account for additional variance in group differences, and (d) there were no significant trends in implementation across time as measured by any of the tools, however there were significant differences in trends over time between groups when using the OTR measure. These findings will be discussed in detail next, however collectively, these results suggest that while these observation tools are correlated, they appear to be capturing different aspects of implementation, and therefore offer support for measuring implementation from multiple perspectives.

Of particular interest in this study was a comparison of both *what* and *how* these tools measured implementation while holding the instructional context constant. As such, close attention was paid to the different components of implementation (i.e., adherence to program theory, dosage, quality of delivery, participant responsiveness) and different measurement approaches (i.e., discrete behavioral observation, global rating scales) of each tool throughout this analysis, and how these measurement features interacted with the context of implementation. These implications will be considered next, after an

overview of the primary findings of this study, and a discussion of the study's limitations. Finally, the chapter will conclude with a discussion on implications for future research and practice.

Primary Findings

Prior to discussing the findings of this study, and their implications for researchers and practitioners, it is important to reiterate the caveat that this study was highly exploratory in nature. This, when coupled with the small sample size and subsequent power issues, limits the certainty with which these patterns can be interpreted and generalized.

A second major caveat has to do with the floor effects that were observed in the student outcome measure. Floor effects may conceal variability between individuals who score at the lower end of a measure, and therefore lessen a measure's ability to detect true individual differences. Floor effects also weaken correlations between the outcome measure and predictors (Catts et al., 2009). Given the fact that over half (n = 18) of the students in the sample scored a zero on the WAT, the results obtained here are *underestimates* of the predictive effects of all three implementation measures. These issues will be discussed in depth in the limitations section, however their impact is such that it is important to preface all discussion and interpretations with this consideration.

Relationship between tools. The three implementation measures were significantly positively and strongly related to each other. In other words, the tools appear to capture highly related elements of instructional implementation – high measures of OTRs align with higher quality instruction on the CLASS, and higher ratings of instructional delivery on the QIDR. More research is needed to examine the true

relationship between these implementation measures to understand how they interact and account for variance beyond looking at basic correlations.

Associations between individual tools and student outcomes. At its core, in educational contexts, implementation science research is undertaken in an effort to better understand and interpret program effects on student outcomes (Fixsen et al., 2005; O'Donnell, 2008). In the current study, the associations between implementation and student outcomes were examined three ways. First, a measure of opportunities to respond was examined, which targeted the frequency of verbal teacher-provided opportunities for students to engage in instructional activities. Second, a composite version of the CLASS that included all ten items across the three subscales was examined (Hamre et al., 2014; McGinty et al., 2012). Finally, a composite of two of the four subscales of the QIDR was examined, which targeted Quality of Intervention Delivery and Student Response During Delivery. In this context, when looking at each implementation measure individually, none of these three tools were statistically significant predictors. However, the examination of model statistics indicated that in the QIDR model, a third of the variance was between groups, and the QIDR, while not significant, accounted for a substantial portion (36%) of that variance on the WAT. The OTR measure and the CLASS were not adequate predictors of group differences in student outcomes.

OTR. The finding that the OTR measure was not a significant predictor of group differences in student outcomes was somewhat surprising. High numbers of OTRs have long been held as important markers of effective instruction, particularly in special education contexts, and there is a growing body of literature (e.g., Doabler et al., in press; Smolkowski & Gunn, 2012) that documents the strength of individual OTRs in particular

in predicting student outcomes for early academic skills. The fact that both overall and individual OTRs were not strong predictors of student outcomes in the current study is therefore unexpected, though there may be several plausible explanations for this finding.

One consideration is the sheer number of OTRs (i.e., an average of 252) that students received both individually and as a group. This high number may be due to the nature of the OTR measure used, which was purposefully liberal in order to capture both social and academic interactions between teachers and students. It's more likely, however, that this high number is due to the nature of the explicit instruction programs used, in which OTRs are a critical element. As such, it's possible that there is a threshold effect at play, where OTRs are predictive of student outcomes up to a certain point, but the sheer number provided by these scripted programs limits any predictive or differentiating power of this measure. This is in line with other implementation research, where some studies have found that with certain elements of implementation (e.g., dosage, adherence), more is not always better, and that there may in fact be a level of diminishing returns once a certain threshold of core components have been delivered (Durlak & DuPre, 2008; Fuchs et al., 2010; Harn et al., 2013).

A second consideration has to do with the above-noted floor effects that were observed in the student outcome measure. While the correlations between all three tools and the WAT increased strongly when only students who performed above the floor were included, the correlation between the OTR and WAT was significant ($r = .79$, $p < .05$), which is notable with a sample of only 13 students. As such, these findings should not discount the importance of OTRs, but rather should stress the importance of examining any possible threshold effects of OTRs, with an adequately sensitive outcome measure.

CLASS. The finding that the CLASS alone did not account for group differences in student outcomes is not necessarily unexpected, though again these results should be considered with in light of the floor effects noted above. While the CLASS has been found to be a strong predictor of student outcomes in other educational contexts (e.g., K-12 general education contexts; Hamre et al., 2007; Mashburn et al., 2008), there is little research that documents this tool's impact in small group or special education settings. The CLASS was developed to capture components of effective instruction in general education classroom contexts, however the structure and content of the programs used in the current study are tailored to support students in need of targeted intervention that is delivered in small, academically homogenous groups. As such, the theories of instructional change involved in the CLASS may not accurately capture the active ingredients that drive student learning in a targeted explicit instruction program.

A discussion of the specific components of the CLASS and how they mapped on to the observed instruction in this study's data set may shed light on this finding. The scores on the CLASS composite for this sample were in the low- mid-range (M =3.56 on scale of 1-7, where scores of 1-2 are considered low, 3-5 are considered medium, and 6-7 are considered high), with much lower overall ratings of instruction than the other two measures. Initial reviews of the specific CLASS components indicated that the classroom organization and management domain was in the high range, but that the domains related to emotional and instructional support were in the low and low-mid range. These patterns were consistent with hypothesized results, and may in fact represent areas of misalignment between program and measure theory (Mowbray et al., 2003; O'Donnell, 2008).

The programs used in this study have strong behavior management and academic supports built in to maximize student engagement and productivity, which may explain the consistently high scores on the Classroom Organization scale. This may also reflect a difference between the tool and program designs, as the CLASS is typically used in large group settings for a longer period of time; quite different than the 30-minute small group setting here. While this setting and these types of programs are designed to provide strong organization and management, the limited variability on the high end of this scale may also indicate that the CLASS is unable to differentiate between overall high levels of classroom organization and management in small group settings. This may indicate a misalignment between the CLASS and the current instructional context, at least in terms of classroom organization and management.

There was, however, more variability between teachers on the Emotional Supports scale, which measures overall classroom climate, teacher-student relationships, and teachers' sensitivity to student needs and perspectives. The presence of some variability on this scale is somewhat surprising, given the scripted nature of the programs, however this may indicate that the CLASS may in fact be able to differentiate between teachers who provided more or less social and emotional supports, even within the context of highly scripted small group instruction.

Uniformly low scores in the area of instructional support, *as defined by the CLASS framework*, while not necessarily surprising, may represent a clear example of a misalignment between program and measure theory (Mowbray et al., 2003; O'Donnell, 2008). The Instructional Support scale of the CLASS measures a teacher's ability to develop higher order thinking skills, to engage students in extended back and forth

feedback loops, and to ask open-ended questions, among other items (Pianta et al., 2008). These skills are clearly important aspects of general instructional quality, but may not align with the context of measurement in the current study. In fact, teachers who engage in the types of instructional supports targeted by the CLASS may actually be considered as "failing to adhere" to the core instructional ingredients of explicit programs.

For example, the programs used here are designed to provide significant instructional support in the development of basic early literacy skills. This involves the provision of carefully sequenced targeted learning opportunities through the use of frequent practice opportunities, delivered at a brisk pace, with clear and consistent language and feedback, particularly error corrections (Archer & Hughes, 2011). This type of specialized instruction is designed to target new skill development with intensive, repetitive direct modeling and immediate corrective feedback to promote errorless learning (Jones & Brownell, 2014), which is directly in contrast to the type of feedback and concept development targeted by the Instructional Support scale. As such, the findings in the current study may indicate that CLASS alone, due to the lack of alignment with explicit instruction theory, is unable to account for the type and quality of instructional support intended for the current measurement context. These results should be examined further, as the different CLASS components could not be examined systematically due to the power limitations in the current study, however the patterns found above with the CLASS are consistent with theoretical hypotheses that indicate that the CLASS alone may not be the best measure of instructional implementation in tier two or special education contexts.

QIDR. Given the stated sample size and floor effect issues, the finding that the QIDR did individually account for substantial variance between groups is noteworthy. Several considerations warrant discussion. Theoretically, this finding was expected, because the QIDR was developed specifically for use with small group targeted interventions, and while it is content-independent, and therefore not specifically aligned with the reading programs used in this study (i.e., a measure of procedural adherence), it is rooted in theories of explicit instruction. Alignment with program theory is considered a particularly important component of implementation measures, particularly for capturing more complex, nuanced features of implementation that are assumed to drive change in student outcomes (Harn et al., 2013; Mowbray et al., 2003). One explanation for the finding that the QIDR accounted for substantial variance between groups may be the fact that this tool aligns with the specific intervention program theory (Mowbray et al., 2003; O'Donnell, 2008).

A second consideration is that the QIDR is a multifaceted measure, as it targets multiple aspects of implementation that are thought to explain important instructional interactions in a small group intervention context. As an integrated tool, the QIDR examines both structural and process-oriented aspects of implementation (i.e., adherence to program theory, quality, student responsiveness). Another plausible explanation for the current study's findings may be that by targeting and capturing multiple components of implementation, the QIDR accounted for multiple instructional interactions that impact student outcomes. This is in line with findings from recent research using multifaceted measurement approaches (e.g., Domitrovich et al., 2010; Durlak & DuPre, 2008; Hamre et al., 2010; Odom et al., 2010). The OTR and CLASS measures captured structural and

process-oriented aspects of implementation, respectively, however these alone may be too narrow to distinguish between important aspects of instructional groups. The QIDR, on the other hand, by addressing both structural and process-oriented aspects, in alignment with program theory, may capture essential elements of instruction that can meaningfully discriminate between groups.

Specifically, the QIDR overtly measures features of student responsiveness as a factor of implementation. The structural components of group responsiveness map on closely with explicit instruction program theory (e.g., systematic, frequent practice opportunities, scaffolding and monitoring of student accuracy), however the QIDR also qualitatively rates students' academic, social, and emotional responsiveness as a group. This may be a particularly important indicator with students who are at-risk for special education, whose group and individual responses should drive instructional decisions (Chard et al., 2002; Connor, 2013; Connor et al., 2009; Harn, Chard, Biancarosa, & Kame'enui, 2011; Jones & Brownell, 2014; Vaughn et al., 2000). Given the at-risk nature of the sample, and its focus on small group intervention, implementation measures that capture more nuanced elements of the dynamic interactions between students and teachers during instructional delivery may be essential in evaluating quality intervention implementation in special education contexts (Harn et al., 2011; Jones & Brownell, 2014). In measuring both the structural and process-oriented aspects of student response to the specific intervention programs, the QIDR may be capturing highly contextual and important information about student responsiveness that differentiates instruction by group. Future research is needed that systematically examines the relationship between the different aspects of implementation within the QIDR, and different student outcomes.

Unique variance in dual predictor models. This study also sought to examine the ways these implementation tools explained variance in student outcomes when entered into models simultaneously. Given the small sample size, only dual predictor models were viable options for examining this research question. The CLASS and OTR measures were poor predictors individually. The OTR and QIDR model increased the variance at level two, and accounted for less of that variance than the QIDR alone, indicating that the OTR simply "added noise" to the model. The CLASS and QIDR model, on the other hand, increased the level two variance explained from 36 to 55%, indicating that together, the CLASS and QIDR accounted for more variance between groups. Closer examination of this model and the correlational analysis, however, indicated that suppression may be occurring between these two predictor variables. This will be discussed in the next section.

QIDR and CLASS. The finding that the CLASS increased the overall variability accounted for in the WAT is not altogether unexpected, but it is again notable given the methodological limitations. Two considerations warrant discussion. The first is the relation between the tools; the second is an extended discussion of the impact of suppression effects.

The CLASS is often considered a general measure of overall classroom quality (e.g., Domitrovich et al., 2010; Hamre et al., 2014; McGinty et al., 2012). There is growing recognition that implementation of any instructional program does not occur in a vacuum outside the influence of good teaching (Durlak, 2010), and recent research has explored relationships between the CLASS as a more general tool and more targeted, implementation measures (i.e., Domitrovich et al., 2010; Hamre et al., 2010; McGinty et

al., 2012). In the current study, the CLASS is not a strong sole predictor of group differences, possibly because it is not capturing adequate levels of intervention-specific variance. However, when used in combination with the QIDR, the CLASS may be capturing enough additional, unique aspects of general teaching quality that it increases the variance accounted for in intervention delivery. In this context, therefore, the CLASS and QIDR may be complementary, as the CLASS captures overall classroom quality, while the QIDR targets specific instructional and program delivery features (i.e., adherence to program theory, quality of instruction, participant responsiveness). This aligns with other studies that show that the CLASS as a measure of general quality targets different information than measures that target the quality of explicit instruction (Domitrovich et al., 2010; McGinty et al., 2012).

These findings must also be considered in terms of suppression effects. Suppression effects occur when a predictor variable that is not highly related to the outcome measure is highly correlated to another predictor; when these predictors are entered into a model simultaneously, the first predictor "suppresses" additional irrelevant variance in the model, therefore increasing the overall predictive power of the model (Pedhazur, 1997). When suppression occurs, illogical parameter estimates (e.g., a measure that should predict increases in an outcome suddenly predict decreases) are also common, which may account for the negative CLASS coefficient seen in Model 7. It is important to acknowledge suppression as an issue in the current study, as this negative CLASS correlation should not be interpreted as meaning that groups with lower quality ratings were associated with higher student outcomes. In fact, neither the CLASS nor

QIDR coefficients should be interpreted independently here; all that can be interpreted is that together they account for more variance than the QIDR alone.

It may be helpful to consider how the different predictors impact group scores in the QIDR versus QIDR and CLASS models. Groups that are rated high, or low, on the QIDR, are also rated comparably on the CLASS. As such, the estimates for individual groups are closer to the obtained values when their scores are combined in the dual predictor model. Another view of this is illustrated by the fact that suppressor variables can also act as "enhancers," in that they subtract irrelevant variance that they share with the other predictor from the model, and therefore improve the variance that is explained (Pedhazer, 1997). It's highly possible that the CLASS is serving in this function in the current study as it contributes to the amount of variance accounted for by the QIDR, which would align with the view that the CLASS, as a general measure of quality, complements more instructionally specific measures of implementation. Further research is necessary, however, to examine the true relationship between these two implementation measures.

Implementation across time. With this final research question, this study also sought to examine the ways these tools captured patterns of implementation across time. These interventions were delivered in schools with well-established RTI frameworks, by IAs with years of experience with the specific explicit instruction programs in use. Theoretically, patterns of implementation across time have been predicted to increase, or decrease, depending on the context of delivery and type of program in use, making this type of inquiry especially important (Durlak, 2010; Zvoch, 2009). In the current study, there were no significant changes across time, on average, with any of the

implementation tools. The OTR measure, however, documented significant between group differences across time. Here again there are several plausible explanations for these findings.

Type of measure. First, these findings should be considered in light of how each of these tools measures implementation. The OTR, as a discrete behavioral observation tool, captures more molecular differences in implementation across time. The CLASS and QIDR, as global rating scales, target more stable elements of implementation, and are therefore expected to capture less variability across time (Chomat-Mooney et al., 2008; Snyder et al., 2006). As such, the finding that the OTR captured significant between group differences across time, while the QIDR and CLASS did not, makes sense given that the OTR measure focuses on "state-like" behaviors that are expected to vary more across time, and is in line with previous research comparing these types of measures (Chomat-Mooney et al., 2008; Cobb & Smith, 2008; Klingner et al., 2013; Snyder et al., 2006; Stoolmiller et al., 2000). This finding is also corroborated by the fact that while the differences attributable to change over time were higher for the OTR measure, it actually explained less between group variance (22%) than the CLASS (44%) and QIDR (74%) when examined across time. This again indicates that the two global rating measures may be capturing more stable, trait-like constructs, while the OTR captures more variable state-dependent elements of implementation.

While these overall results are expected, it was somewhat surprising that the QIDR did not capture more between group differences across time, given that it integrates more behavioral elements into both what and how it measures implementation. Closer examination of the patterns of implementation across time as measured by the

QIDR show, however, that while there were no significant upward or downward trends (i.e., growth) within groups, there was a substantial degree of instability within groups, indicating that this measure may be capturing a wider range of behaviors within a general trait-like level of implementation. Figure 7 illustrates the variability of implementation within group. Additional research is needed to examine the integrated nature of the QIDR, in terms of whether it is molecular enough to capture state-like differences that can inform formative professional development and coaching, while also capturing more stable trait-like differences in small group instructional delivery.

Context. These findings should also be considered in light of the context in which these interventions were delivered. Both schools were considered to be in the full operation stage of implementation with their RTI program and kindergarten interventions. This is important for several reasons, as it not only means that a certain level of competency, or expertise, was expected from the IAs delivering the programs, but it also meant that no professional development or coaching was provided. The fact that there were no significant trends overall, then, makes sense for the given context, and these findings can perhaps best be thought of as representing a snapshot of implementation across time during the fully operational implementation stage. These findings support the argument that different stages of implementation may require different measurement approaches (Fixsen et al., 2005), such that interventions in later implementation stages may need less observation across time than those in less stable earlier stages.

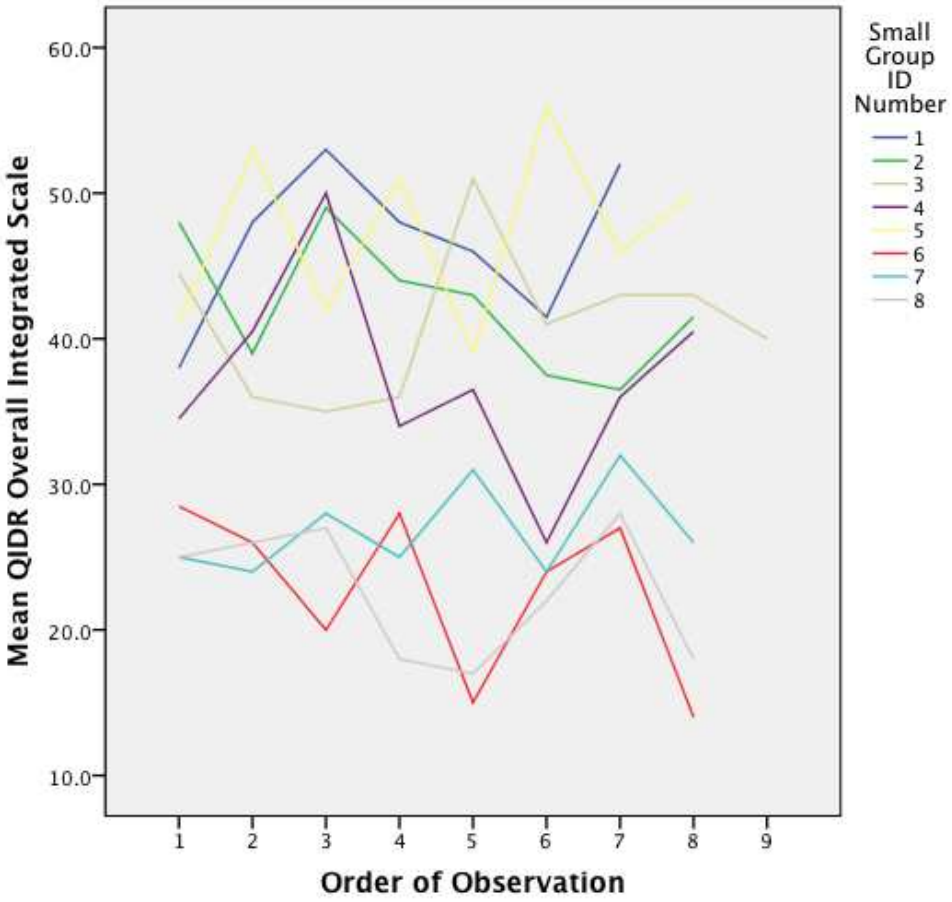

Figure 7. Variability of QIDR implementation across time.

Further research is needed to examine these patterns however, as the between group differences found with the OTR, the instability seen in the QIDR, and the relative stability seen with the CLASS may suggest that measures that target more variable behaviors require different frequencies of observations across time, even within the fully operational implementation stage. For instance, the instability of the QIDR measure may suggest that taking an average score across weekly observations gives a more accurate measure of overall implementation. The relative stability of the CLASS may indicate that fewer observations across time are necessary. For the OTR, on the other hand, the significant differences in growth between groups may indicate that an average score masks important variability. Students in groups with significant downward trends may be

more impacted by later low scores; students in groups with significant upward trends may not have benefited from all the instructional supports needed across time. Decisions around the timing of OTR observations may therefore need to be group or context-specific, rather then set at regular intervals across time.

These findings align with the findings of Chomat-Mooney et al. (2008), which found that only four observations were required to provide a relatively stable measure of classroom processes when using the CLASS, whereas at least six observations were required when using a behavioral, time-sampling approach. Further research is necessary to examine the stability of the QIDR, particularly given its integrated approach, to determine the ideal number of observations needed to provide a reliable, stable measure of implementation. Scale differences should also be examined, as the differences between these measures, where the CLASS had a range of 2.25-5.00, and the OTR measure had a range 116-458, may also impact how these measures capture variability across time.

Basic adherence. Finally, an additional noteworthy finding of this study is the lack of strong correlations between the basic procedural adherence measure and the implementation tools. All teachers across time had high levels of procedural adherence (M = 5.63 on a scale of 1-6), indicating that as a group they were delivering the programs "by the book," as would be expected from schools in the fully operational stage of implementation. Yet this measure was not strongly associated with the other implementation measures (r = .36 with the OTR; r = .22 with the CLASS, r = .35 with the QIDR), indicating that basic adherence to procedure alone may not be a sufficient measure of instructional implementation. This is an important finding given the

widespread approach of only including basic adherence measures when evaluating implementation, particularly in education contexts (Harn et al., 2013; O'Donnell, 2008).

Several measurement issues may have led to this finding, such as the lack of variability on the adherence measure, or the scaling of the measure. The dichotomous nature of this tool, where teachers' basic adherence to procedures was rated as either present or not, may have decreased variability and impacted the correlations. Future research is needed to more fully examine the relationship of measures of basic adherence to procedural and other more multifaceted measures of implementation, however the findings of the current study indicate that commonly-used basic adherence measures may be inappropriate for capturing the full complexity of instructional implementation in school-based settings.

Limitations

There are several limitations to the current study that warrant consideration, which also may be helpful in informing the design of future studies. First, it is important to underscore the underpowered nature of this study. While the current study's sample represented naturally occurring clustering in a school-based setting, the final sample of 31 students nested in only 8 small groups was problematic from a sample size perspective. Studies with insufficient power increase the likelihood of type II errors, and lead to inconsistent identification of statistically significant effects (Maxwell, 2004). These power issues also limited the model building process, such that no student-level predictors could be entered into the model, and entering all three tools into the model simultaneously resulted in an oversaturated model. This limited the range of questions that could be addressed in the current study (Snijders, 2005; Snijders & Bosker, 1999).

The fact that none of the tools were significant predictors is also likely due to the underpowered nature of this study, though the fact that substantial variance was still accounted for by some of the models indicates that that a better-powered study may be able to detect significant effects using these implementation measures.

A second limitation had to do with the outcome measure in the current study. The WAT was not sensitive enough to detect individual differences in outcomes in the current sample, as evidenced by serious floor effect issues. As over half of the sample scored a zero, these floor effects led to what is most likely an underestimation of the predictive power of the three implementation measures. The WAT was selected as an outcome measure as it targeted the type of early literacy skills being taught in the interventions, however it is highly possible that a more sensitive measure (e.g., DIBELS, easy-CBM) may have captured more accurate differences in a sample made up of students likely to score toward the lower end of a distribution.

An additional limitation has to do with the high level of collinearity among the three implementation tools, which was particularly problematic given the power and floor effects noted above. Instructional implementation appears to be highly consistent across all three measures, where groups who score highly on the CLASS and QIDR also offer higher numbers of OTRs. This led to the likely occurrence of suppression effects, which limits the interpretability of the coefficients in the models. However the finding that the QIDR and CLASS do account for more variance when modeled simultaneously may indicate that a better powered study may detect the unique variance attributed by these highly related tools. Collectively, these limitations indicate that the current study represents a lower bound of the explanatory effects of all three measures, and future

research is needed to determine the relationships and effects of implementation as measured by these tools. Despite these limitations, the current study does offer important insights to the field in terms of measuring implementation, which are discussed next.

Implications for Research

From a researcher perspective, a comprehensive view of implementation involves measuring implementation as a complex, multifaceted construct (Dane & Schneider, 1998; Durlak & DuPre, 2008; Fixsen et al., 2005; Harn et al., 2013; Mowbray et al., 2003; O'Donnell, 2008). This type of measurement approach must be rooted in a firm understanding of the *purpose* of implementation measurement. Clear consideration of questions about the purpose of implementation measurement involves systematic decisions about (a) what to measure, (b) when to measure it, and (c) how to measure implementation variables. The findings of the current study offer strong support for considering implementation from this view, and the need to take a multifaceted approach to measuring implementation.

Implications for what to measure. The results of this study challenge the typical implementation measurement approaches often used in special education contexts that only address unidimensional measures of adherence or fidelity to procedure. While adherence to basic program procedures was found to be high across the entire sample, it was not associated with other measures of implementation, and these other measures appear to be accounting for important differences between small groups. The finding that the QIDR was the only tool to account for substantial differences between groups reinforces the notion that implementation measures should also be closely aligned with the program theory, content, and context of delivery. When researchers select measures

based on a firm understanding of these underlying intervention mechanisms, rather than simply following convention or ensuring the internal validity of a research study, implementation measures are more likely to capture important variability in delivery (Durlak, 2010; Flay et al., 2005; Mowbray et al., 2003; O'Donnell, 2008).

These results also align with recommendations calling for the use of multiple measures that target different aspects of implementation (Dane & Schneider, 1998; Durlak & DuPre, 2008; Odom, Hansen et al., 2010), in that the QIDR and the CLASS together accounted for more variance between groups than the QIDR alone. These findings are supported by research showing that both general classroom quality and content-specific instructional quality may be important constructs to capture (e.g., Domitrovich et al., 2010; Hamre et al., 2010; McGinty et al., 2012) when developing or selecting implementation measures. This finding, despite the extremely high collinearity between the process-oriented CLASS and the integrated QIDR, also supports the need to take a multifaceted approach to measuring implementation. Researchers need to consider measures that target both the structural and process-oriented aspects of implementation (e.g., Odom et al., 2010), and may want to consider the development of multifaceted composites that capture a wider range of implementation variables, though a systematic examination of an OTR/CLASS/QIDR composite was beyond the scope of the current project.

Implications for when to measure implementation. The current study also contributes to implementation science research by examining the dynamic nature of implementation across time (e.g., Chomat-Mooney et al., 2008; Domitrovich et al., 2010; Fixsen et al., 2005; Zvoch, 2009). While research in this area is still emerging, with many

equivocal results, the current findings, which indicate that overall implementation was stable across time, may best be considered as a snapshot of implementation in a given implementation stage and context that can be used to inform future researchers. In general, it may be expected that schools that are far along in the implementation stages, with well-established practices and systems of support, may have fairly stable levels of implementation across time, and as such may not require as many observations in order to establish accurate measures of implementation. These findings align with theories about implementation stages (i.e., Fixsen et al., 2005), and support the call for researchers to take the context and stage of implementation into consideration when making implementation measurement decisions.

Closer examination of the OTR across time results, however, indicate that the significant between group differences may be tied to broader contextual factors (i.e., school site), as three out of the four groups with downward trends of OTR were located at the same school. While both schools were part of the same district, received similar trainings, and were in the fully operational stage of RTI implementation, there were other between-school differences that may have impacted implementation across time. Given these findings, a closer examination of overall school differences indicated that School A, which had smaller groups, instruction delivered in a common room, and consistent supervision by a lead teacher, had average WAT scores of 105.69, while School B, where students received instruction in larger groups in isolated rooms, had average WAT scores of 95.72. While systematic examination of these factors was beyond the scope of the current study, a review of these types of contextual variables should also be included

when making implementation measurement decisions. If implementation had only been measured at one or even three points in time, these patterns may not have been noticed.

Indeed, the finding that overall implementation across time was relatively stable but that individual differences were noted with different tools, raises the question of how many observations are necessary to capture a valid picture of implementation. These patterns also suggest that this question may not only be a function of context, but the type of measure. The relative instability of the QIDR suggests that taking an average of multiple observations may be necessary for tools that target more variable components of implementation, while the significant between group differences on the OTR suggest that multiple observations should be examined across time and contexts to capture important moment to moment differences. The relative stability of the CLASS, on the other hand, aligns with recommendations from the CLASS literature (Chomat-Mooney et al., 2008; Pianta et al., 2008) that only a few observations may be enough, even in this small group setting. Collectively, these results strongly support the call to move beyond single measures of implementation to a more dynamic, theoretically aligned measurement approach across time (Durlak & DuPre, 2008; Fixsen et al., 2005; Harn et al., 2013; Mowbray et al., 2003; Zvoch, 2009).

Implications for how to measure implementation. The differences in how global versus behavioral measurement approaches capture implementation across time has important implications for researchers, and is in line with earlier research highlighting the difference between measures that capture state-like versus trait-like behaviors (Chomat-Mooney et al., 2008; Snyder et al., 2006). Researchers need to carefully consider the purpose of measurement (e.g., formative versus summative

evaluation) when selecting or developing implementation measures, and consider how these align and capture information about the type of instructional processes being observed. For instance, when evaluating program use and effectiveness, different programs may require different combinations of tools targeting different levels of instructional information (i.e., instructional states that changes across time and contexts, or more stable instructional traits), such as a special education context that requires discrete information on the dosage of specific instructional interactions individual students need, as well as overall information about classroom quality and other process variables (e.g., Connor et al., 2013; Doabler et al., in press; McGinty et al., 2012). Regardless of which measurement approach is used, these findings collectively underscore the importance of aligning implementation purpose and implementation measures.

Implications for linking implementation and outcomes. Finally, the findings in the current study collectively reinforce the call to examine implementation variables in terms of the role they play in mediating or moderating student outcomes. While none of the tools used here were significant predictors of student outcomes, due to the underpowered nature of the study, the substantial variance explained between groups by the QIDR, and the QIDR and CLASS together appear to indicate that these measures are capturing important group differences that may impact student outcomes in a larger sample. The finding that the QIDR and CLASS together accounted for more variance also supports the need to take a multifaceted view when understanding the role instructional implementation plays in impacting student outcomes, as together these

highly collinear tools appear to shed light on classroom processes that account for important group differences.

Implications for Practice

With the understanding that teachers play an important role in shaping student achievement, more and more research and policy resources are being allocated to the evaluation of instruction and the "black box" of classroom processes (Connor, 2013; Jones & Brownell, 2014; Pianta & Hamre, 2009). More direct observation measures are being developed, and adopted by states and districts for the use of teacher evaluation, and in turn for use in making high-stakes decisions about teacher effectiveness. While the current study is aimed at supporting researchers in their understanding, development, and use of implementation tools, findings also hold important implications for practitioners.

First, these findings highlight the importance of considering context when measuring implementation. The findings that the CLASS and OTR measures alone did not account for substantial variance in student outcomes in tier two small group intervention settings is worth noting, as both are widely used implementation/"fidelity" measures. As more direct observation tools are used to measure teacher and instructional quality, examining the differences between general and special education contexts, instructional approaches, and evidence-based practices will become more important. So too, will the development of tools that align with the evidence-based programs used in these settings that accurately evaluate the wide array of specialized instruction that is necessary to support students with special needs (Connor, 2013; Jones & Brownell, 2014). The QIDR, while in need of further validation, may be one such tool that can validly assess instruction in targeted tier two small group settings. It is important to

caution, however, that relying on any one tool, even one that's developed to incorporate multifaceted aspects of implementation, may be problematic, particularly given the increasing complexity and variety of instructional roles special education teachers are asked to fill. The fact that the QIDR and CLASS together explained more variance than the QIDR alone underscores this issue. Researchers and practitioners should work together to determine which measures, or combinations of measures, are best able to accurately and validly evaluate the instruction that's provided across given contexts.

A second consideration for practitioners is whether and how implementation measures can be used to support professional development. Some elements of instruction may require very specific levels of feedback (e.g., dosage of specific behaviors), whereas others may require more nuanced feedback and support (e.g., quality of instructional delivery). The different types of measures examined here (i.e., global vs. behavioral) may therefore be helpful for different evaluation purposes (Chomat-Mooney et al., 2008; Doabler et al., in press; Snyder et al., 2006). Formative teacher evaluations, for example, may require more behavioral measures like the OTR that can be used to target modifiable behaviors, whereas summative evaluations that are targeting more stable trait-like differences between teachers or classrooms may require the use of global measures. These global rating tools may also be used to shape conversations and larger professional development efforts around unified views of effective instruction at the school and district level (Connor, 2013; Raudenbush, 2009). The QIDR, as an integrated tool, may perhaps be useful across both formative and summative professional development contexts. Future research is needed, however, to determine the levels and types of

training that are necessary to prepare typical school personnel to reliably, validly and efficiently use various implementation measures.

Future Research

As previously mentioned, the results of the current study are exploratory, and future research is needed that addresses the power and sample limitations of this study. Further studies are also needed to examine models that incorporate theoretically important student-level variables (e.g., initial student skill level, special education or ELL status), and to address the floor issues in the current study. Studies with additional outcome measures (e.g., more sensitive curriculum-based early literacy measures) may shed further light on the relationship between these measures of implementation and outcomes, though additional measures (e.g., measures of self-regulation, language outcomes, behavioral skills) may also provide important insight into the relationship between implementation and student achievement. Future studies may explore whether the different implementation measures examined here differentially predict different outcomes, as there is growing support for this type of pattern across the literature base (e.g., Domitrovich et al., 2010; Hamre et al., 2014; Odom et al., 2010).

Additional research with a larger sample size may also shed light on the relationship between the CLASS and the QIDR, given the suppression effects in the current study. A study examining composites of the implementation measures may also provide more information about the predictive power of these tools. Additional studies that compare how these tools measure instruction across different contexts (e.g., small groups within general education classrooms versus small groups in pullout/resource room settings) is also warranted, and may detect further differences between tools.

Additional research into training observers in the use of these observation tools is also warranted. In the current study, the OTR measure required the most training and support, and the CLASS and QIDR required significantly less training than is typically reported with global measures. It is unknown if this has to do with the observers who worked with these tools (i.e., highly trained graduate students), or the instruction observed (i.e., video tapes of small group instruction versus large classroom settings), or some combination of these and other factors. Finally, further research is needed to examine how these tools measure implementation across time. The results here may suggest that each tool requires a different number of observations to predict stable patterns of implementation.

Conclusions

Implementation is a complex task, dependent on content, context, timing, and how it is conceptualized. Implementation measurement must therefore be an equally complex process, particularly as a means of understanding how instruction and classroom processes impact student outcomes. Findings from the current study offered initial support for the use of a multifaceted approach to measuring implementation, and found that the QIDR, as an integrated, theoretically-aligned measure of small group explicit interventions, accounted for substantial variance between small groups, particularly when coupled with the CLASS, a more general, process-oriented measure of classroom quality. These results also indicate that while implementation across intervention time was stable overall, the OTR, as a more molecular view of teaching behaviors, was able to detect significant differences in groups across time. While additional research is needed to examine the relationships between these measures and student outcomes, the current

study offers important exploratory insights into the complex process of measuring implementation in school-based settings.

APPENDIX A

OTR OVERVIEW

Opportunities to Respond (OTRs): Any teacher-provided (**prompted**) opportunity for students to answer questions, practice content, read aloud, or actively participate in instructional tasks.

Group OTRs: Teacher provides an opportunity for students to respond as a group (*choral responses, written responses, thumbs up signal*)

Individual OTRS: Teacher provides individual students an opportunity to respond (*round robin turns, individual practice, individual questions**)***

Defer to Individual OTRs

Types of OTRs:

- Oral Responses
  - Choral Responding
  - Partner Responding
  - Team Responding
  - Individual Responding
- Written Responses
  - Response Slates
  - Response Cards
  - Work Sheets
  - White Boards
- Action Responses
  - Hand Signals
  - Gestures

Teacher Feedback: Any teacher-provided response (**verbal**) to student responses or student behavior.

Praise: Any positive verbal teacher responses to student behaviors *("Good work," "Nice job raising your hand,"* or *"Yes, that word is man"***).**

Corrective Feedback: Any feedback that indicates that an academic or behavioral error occurred *("That letter says sssss," "That word is pat," "Pencils on desks, not in your hands please").*

OTR Recording Sheet

Coder Name: _____ Video ID:

Time	OTRS		Teacher Feedback	
	Group OTR	Individual OTR	Praise	Corrective Feedback
0:00-0:30				
0:30-1:00				
1:00-1:30				
1:30-2:00				
2:00-2:30				
2:30-3:00				
3:00-3:30				
3:30-4:00				
4:00-4:30				
4:30-5:00				
BREAK	Total:	Total:	Total:	Total:
5:00-5:30				
5:30-6:00				
6:00-6:30				
6:30-7:00				
7:00-7:30				
7:30-8:00				
8:00-8:30				
8:30-9:00				
9:00-9:30				
9:30-10:00				
BREAK	Total:	Total:	Total:	Total:

Time	OTRS		Teacher Feedback	
	Group OTR	Individual OTR	Praise	Corrective Feedback
10:00-10:30				
10:30-11:00				
11:00-11:30				
11:30-12:00				
12:00-12:30				
12:30-13:00				
13:00-13:30				
13:30-14:00				
14:00-14:30				
14:30-15:00				
BREAK	Total:	Total:	Total:	Total:
15:00-15:30				
15:30-16:00				
16:00-16:30				
16:30-17:00				
17:00-17:30				
17:30-18:00				
18:00-18:30				
18:30-19:00				
19:00-19:30				
19:30-20:00				
BREAK	Total:	Total:	Total:	Total:

Time	OTRS		Teacher Feedback	
	Group OTR	Individual OTR	Praise	Corrective Feedback
20:00-20:30				
20:30-21:00				
21:00-21:30				
21:30-22:00				
22:00-22:30				
22:30-23:00				
23:00-23:30				
23:30-24:00				
24:00-24:30				
24:30-25:00				
BREAK	Total:	Total:	Total:	Total:
25:00-25:30				
25:30-26:00				
26:00-26:30				
26:30-27:00				
27:00-27:30				
27:30-28:00				
28:00-28:30				
28:30-29:00				
29:00-29:30				
29:30-30:00				
BREAK	Total:	Total:	Total:	Total:

APPENDIX B

CLASS OVERVIEW

Positive Climate

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Relationships • Physical proximity • Shared activities • Peer assistance • Matched affect • Social conversation	There are few, if any, indications that the teacher and students enjoy warn, supportive relationships with one another.	There are some indications that the teacher and students enjoy warm, supportive relationships with one another.	There are many indications that the teacher and students enjoy warm, supportive relationships with one another.
Positive Affect • Smiling • Laughter • Enthusiasm	There are no or few displays of positive affect by the teacher and/or students.	There are sometimes displays of positive affect by the teacher and/or students.	There are frequent displays of positive affect by the teacher and/or students.
Positive Communication • Verbal affection • Physical affection • Positive expectations	There are rarely positive communications, verbal or physical, among teachers and students.	There are sometimes positive communications, verbal or physical, among teachers and students.	There are frequently positive communications, verbal or physical, among teachers and students.
Respect • Eye contact • Warm, calm voice • Respectful language • Cooperation and/or sharing	The teacher and students rarely, if ever, demonstrate respect for one another.	The teacher and students sometimes demonstrate respect for one another.	The teacher and students consistently demonstrate respect for one another.

Negative Climate

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Negative Affect • Irritability • Anger • Harsh voice • Peer aggression • Disconnected or escalating negativity	The teacher and students do not display strong negative affect and only rarely, if ever, display mild negativity.	The classroom is characterized by mild displays of irritability, anger, or other negative affect by the teacher and/or the students.	The classroom is characterized by consistent displays of irritability, anger, or other negative affect by the teacher and/or the students.
Punitive Control • Yelling • Threats • Physical Control • Harsh Punishment	The teacher does not yell or make threats to establish control.	The teacher occasionally uses expressed negativity such as threats or yelling to establish control.	The teacher repeatedly yells at students or makes threats to establish control.

Sarcasm/Disrespect • Sarcastic voice/statement • Teasing • Humiliation	The teacher and students are not sarcastic or disrespectful.	The teacher and/or students are occasionally sarcastic or disrespectful.	The teacher and/or students are repeatedly sarcastic or disrespectful.
Severe Negativity • Victimization • Bullying • Physical punishment	There are no instances of severe negativity between the teacher and students.	There are no instances of severe negativity between the teacher and students.	There are instances of severe negativity between the teacher and students or among the students.

Teacher Sensitivity

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Awareness • Anticipates problems and plans appropriately • Notices lack of understanding and/or difficulties	The teacher consistently fails to be aware of students who need extra support, assistance, or attention.	The teacher is sometimes aware of students who need extra support, assistance, or attention.	The teacher is consistently aware of students who need extra support, assistance, or attention.
Responsiveness • Acknowledges emotions • Provides comfort and assistance • Provides individualized support	The teacher is unresponsive to or dismissive of students and provides the same level of assistance to all students, regardless of their individual needs.	The teacher is responsive to students sometimes but at other times is more dismissive or unresponsive, matching her support to the needs and abilities of some students but not others.	The teacher is consistently responsive to students and matches her support to their needs and abilities.
Addresses Problems • Helps in effective and timely manner • Helps resolve problems	The teacher is ineffective at addressing students' problems and concerns.	The teacher is sometimes effective at addressing students' problems and concerns.	The teacher is consistently effective at addressing students' problems and concerns.
Student Comfort • Seeks support and guidance • Freely participate • Takes risks	The students rarely seek support, share their ideas with, or respond to questions from the teacher.	The students sometimes seek support, share their ideas with, or respond to questions from the teacher.	The students seem comfortable seeking support, sharing their ideas with, and responding freely to the teacher.

Regard for Student Perspectives

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Flexibility and Student Focus • Shows flexibility • Incorporates students' ideas • Follows lead	The teacher is rigid, inflexible, and controlling in his plans and/or rarely goes along with students' ideas; most classroom activities are teacher-driven.	The teacher may follow the students' lead during some periods and be more controlling during others.	The teacher is flexible in his plans, goes along with students' ideas, and organizes instruction around students' interests.
Support for Autonomy and Leadership • Allows choice • Allows students to lead lessons • Gives students responsibilities	The teacher does not support student autonomy and leadership.	The teacher sometimes provides support for student autonomy and leadership but at other times fails to do so.	The teacher provides consistent support for student autonomy and leadership.
Student Expression • Encourages student talk • Elicits ideas and/or perspectives	There are few opportunities for student talk and expression.	There are periods during which there is a lot of student talk and expression but other times when teacher talk predominates.	There are many opportunities for student talk and expression.
Restriction of Movement • Allows movement • Is not rigid	The teacher is highly controlling of students' movement and placement during activities.	The teacher is somewhat controlling of students' movement and placement during activities.	Students have freedom of movement and placement during activities.

Behavior Management

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Clear Behavior Expectations • Clear expectations • Consistency • Clarity of rules	Rules and expectations are absent, unclear, or inconsistently enforced.	Rules and expectations may be state clearly but are inconsistently enforced.	There are many indications that the teacher and students enjoy warm, supportive relationships with one another.
Proactive • Anticipates of problem behavior or escalation • Low reactivity • Monitors	The teacher is reactive, and monitoring is absent or ineffective.	The teacher uses a mix of proactive and reactive responses; sometimes she monitors and reacts to early indicators of behavior problems but other times misses/ignores them.	The teacher is consistently proactive and monitors the classroom effectively to prevent problems from developing.

Redirection of Misbehavior • Effective reduction of misbehavior • Uses subtle cues to redirect • Efficient redirection	Attempts to redirect misbehavior are ineffective; the teacher rarely focuses on positives or uses subtle cues. As a result, misbehavior continues and/or escalates and takes time away from learning.	Some of the teacher's attempts to redirect misbehavior are effective, particularly when he or she focuses on positives and uses subtle cues. As a result, misbehavior rarely continues, escalates, or takes time away from learning.	The teacher effectively redirects misbehavior by focusing on positives and making use of subtle cues. Behavior management does not take time away from learning.
Student Behavior • Frequent compliance • Little aggression and defiance	There are frequent instances of misbehavior in the classroom.	There are periodic instances of misbehavior in the classroom.	There are few, if any, instances of misbehavior in the classroom.

Productivity

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Maximizing Learning Time • Provision of activities • Choice when finished • Few disruptions • Effective completion of managerial tasks • Pacing	Few, if any, activities are provided for students, and an excessive amount of time is spent addressing disruptions and completing managerial tasks.	The teacher provides activities for the students most of the time, but some learning time is lost in dealing with disruptions and the completion of managerial tasks.	The teacher provides activities for the students and deals efficiently with disruptions and managerial tasks.
Routines • Students know what to do • Clear instructions • Little wandering	The classroom routines are unclear; most students do not know what is expected of them.	There is some evidence of classroom routines that allow everyone to know what is expected of them.	The classroom resembles a "well-oiled machine"; everybody knows what is expected of them and how to go about doing it.
Transitions • Brief • Explicit follow-through • Learning opportunities within	Transitions are too long, too frequent, and/or inefficient.	Transitions sometimes take too long or are too frequent and inefficient.	Transitions are quick and efficient.
Preparation • Materials ready and accessible • Knows lessons	The teacher does not have activities prepared and ready for the students.	The teacher is mostly prepared for activities but takes some time away from instruction to take care of last-minute preparations.	The teacher is fully prepared for activities and lessons.

Instructional Learning Formats

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Effective Facilitation • Teacher involvement • Effective questioning • Expanding children's involvement	The teacher does not actively facilitate activities and lessons to encourage students' interest and expanded involvement.	At times, the teacher actively facilitates activities and lessons to encourage students' interest and expanded involvement, but at other times she merely provides activities for the students.	The teacher actively facilitates activities and lessons to encourage students' interest and expanded involvement.
Variety of Modalities and Materials • Range of auditory, visual, and movement opportunities • Interesting and creative materials • Hands-on opportunities	The teacher does not use a variety of modalities or materials to gain students' interest and participation during activities and lessons.	The teacher is inconsistent in her use of a variety of modalities or materials to gain students' interest and participation during activities and lessons.	The teacher uses a variety of modalities including auditory, visual, and movement and uses a variety of materials to gain students' interest and participation during activities and lessons.
Student Interest • Active participation • Listening • Focused attention	The students do not appear interested and/or involved in the lesson or activities.	Students may be engaged and/or interested for periods of time, but at other times their interest wanes and they are not involved in the activity or lesson.	Students are consistently interested and involved in activities and lessons.
Clarity of Learning Objectives • Advanced organizers • Summaries • Reorientation statements	The teacher makes no attempt to or is unsuccessful at orienting and guiding students toward learning objectives.	The teacher orients students somewhat to learning objectives, or the learning objectives may be clear during some periods but less so during others.	The teacher effectively focuses students' attention toward learning objectives and/or the purposes of the lesson.

Concept Development

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Analysis and Reasoning • Why and/or how questions • Problem solving • Prediction/ experimentation • Classification/ comparison • Evaluation	The teacher rarely uses discussions and activities that encourage analysis and reasoning.	The teacher occasionally uses discussions and activities that encourage analysis and reasoning.	The teacher often uses discussions and activities that encourage analysis and reasoning.

Creating • Brainstorming • Planning • Producing	The teacher rarely provides opportunities for students to be creative and/or generate their own ideas and products.	The teacher sometimes provides opportunities for students to be creative and/or generate their own ideas and products.	The teacher often provides opportunities for students to be creative and/or generate their own ideas and products.
Integration • Connect concepts • Integrates with previous knowledge	Concepts and activities are presented independent of one another, and students are not asked to apply previous learning.	The teacher sometimes links concepts and activities to one another and to previous learning.	The teacher consistently links concepts and activities to one another and to previous learning
Connections to the Real World • Real world applications • Related to students' lives	The teacher does not relate concepts to the students' actual lives.	The teacher makes some attempts to relate concepts to the students' actual lives.	The teacher consistently relates concepts to the students' actual lives.

Quality of Feedback

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Scaffolding • Hints • Assistance	The teacher rarely provides scaffolding to students but rather dismisses responses or actions as incorrect or ignores problems in understanding.	The teacher occasionally provides scaffolding to students but at other times simply dismisses responses as incorrect or ignores problems in understanding.	The teacher often scaffolds for students who are having a hard time understanding a concept, answering a question, or completing an activity.
Feedback Loops • Back-and-forth exchange • Persistence by teacher • Follow-up questions	The teacher gives only perfunctory feedback to students.	There are occasional feedback loops – back-and-forth exchanges – between the teacher and students; other times, however, feedback is more perfunctory.	There are frequent feedback loops – back-and-forth exchanges – between the teacher and students.
Prompting Thought Processes • Asks students to explain thinking • Queries responses and actions	The teacher rarely queries the students or prompts students to explain their thinking and rationale for responses and actions.	The teacher occasionally queries the students or prompts students to explain their thinking and rationale for responses and actions.	The teacher often queries the students or prompts students to explain their thinking and rationale for responses and actions.

Providing Information • Expansion • Clarification • Specific feedback	The teacher rarely provides additional information to expand on the students' understanding or actions.	The teacher occasionally queries the students or prompts students to explain their thinking and rationale for responses and actions.	The teacher often queries the students or prompts students to explain their thinking and rationale for responses and actions.
Encouragement and Affirmation • Recognition • Reinforcement • Student persistence	The teacher rarely offers encouragement of students' efforts that increases students' involvement and persistence.	The teacher occasionally offers encouragement of students' efforts that increases students' involvement and persistence.	The teacher often offers encouragement of students' efforts that increases students' involvement and persistence.

Language Modeling

	Low (1, 2)	Middle (3, 4, 5)	High (6, 7)
Frequent Conversations • Back-and-forth exchanges • Contingent responding • Peer conversations	There are few if any conversations in the classroom.	There are limited conversations in the classroom.	There are frequent conversations in the classroom.
Open-Ended Questions • Questions require more than a one-word response • Students respond	The majority of the teacher's questions are closed-ended.	The teacher asks a mix of closed-ended and open-ended questions.	The teacher asks many open-ended questions.
Repetition and Extension • Repeats • Extends/elaborates	The teacher rarely, if ever, repeats or extends the students' responses.	The teacher sometimes repeats or extends the students' responses.	The teacher often repeats or extends the students' responses.
Self and Parallel Talk • Maps own actions with language • Maps student action with language	The teacher rarely maps his or her own actions and the students' actions through language and description.	The teacher occasionally maps his or her own actions and the students' actions through language and description.	The teacher consistently maps his or her own actions and the students' actions through language and description.
Advanced Language • Variety of words • Connected to familiar words and/or ideas	The teacher does not use advanced language with students.	The teacher sometimes uses advanced language with students.	The teacher often uses advanced language with students.

CLASS Observation Sheet

Coder Name: _____ Video ID: _____

Circle appropriate score.

Positive Climate (PC) • Relationships • Positive Affect • Positive Communication • Respect	*Notes*	1	2	3	4	5	6	7
Negative Climate (NC) • Negative Affect • Punitive Control • Sarcasm/Disrespect • Severe Negativity	*Notes*	1	2	3	4	5	6	7
Teacher Sensitivity (TS) • Awareness • Responsiveness • Addresses Problems • Student Comfort	*Notes*	1	2	3	4	5	6	7
Regard for Student Perspectives (RSP) • Flexibility and Student Focus • Support for Autonomy and Leadership • Student Expression • Restriction of Movement	*Notes*	1	2	3	4	5	6	7
Behavior Management (BM) • Clear Behavior Expectations • Proactive • Redirection of Misbehavior • Student Behavior	*Notes*	1	2	3	4	5	6	7
Productivity (PD) • Maximizing Learning Time • Routines • Transitions • Preparation	*Notes*	1	2	3	4	5	6	7
Instructional Learning Formats (ILF) • Effective Facilitation • Variety of Modalities and Materials • Student Interest • Clarity of Learning Objectives	*Notes*	1	2	3	4	5	6	7
Concept Development (CD) • Analysis and Reasoning • Creating • Integration • Connections to the Real World	*Notes*	1	2	3	4	5	6	7
Quality of Feedback (QF) • Scaffolding • Feedback Loops • Prompting Thought Processes • Providing Information • Encouragement and Affirmation	*Notes*	1	2	3	4	5	6	7
Language Modeling • Frequent Conversation • Open-Ended Questions • Repetition and Extension • Self- and Parallel Talk • Advanced Language	*Notes*	1	2	3	4	5	6	7

(Adapted from Pianta, La Paro, & Hamre, 2008)

APPENDIX C

QIDR OVERVIEW

Quality of Intervention Delivery

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).

Item	Not implemented: **0 points** <50%	Inconsistent implementation: **1 point** >50%	Effective implementation: **2 points** >80%	Expert implementation: **3 points** >95%
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.
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.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
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.
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.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
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.
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.

Item	Not implemented: 0 points <50%	Inconsistent implementation: 1 point >50%	Effective implementation: 2 points >80%	Expert implementation: 3 points >95%
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

Examining the Quality of Intervention Delivery and Receipt (QIDR) Recording Sheet

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 is 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	

Examining the Quality of Intervention Delivery and Receipt (QIDR) Recording Sheet (Continued)

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).

APPENDIX D

BASIC PROCEDURAL ADHERENCE OBSERVATION RECORDING SHEET

Coder Name: _____ Video ID: _____

DI Component	Present >80% of the Time?			Comments
	Y	N	N/A	
Teacher has materials prepped and ready for use across the lesson.				
Teacher uses clear, consistent language across the lesson.				
Teacher provides frequent practice opportunities to all students across the lesson.				
Teacher provides clear, consistent error corrections when students make errors across the lesson.				
Teacher delivers instruction at a quick, brisk pace across the lesson.				
Teacher uses clear, consistent signals to alert students to opportunities to respond across the lesson.				

Comments:

REFERENCES CITED

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

Bambara, L. M., Nonnemacher, S., & Kern, L. (2009). Sustaining school-based individualized positive behavior support: Perceived barriers and enablers. *Journal of Positive Behavior Interventions, 11*, 161-176. doi:10.1177/1098300708330878

Beil, S. (n.d.) Early Reading Intervention (ERI) Fidelity checklist. Unpublished instrument. Retrieved from http://www.iup.edu/page.aspx?id=134567

Brophy, J., & Good, T. L.(1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching*, 328-375. New York: Macmillan.

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2009). *Direct instruction reading* (5th ed.). Canada: Pearson Education.

Chard, D. J., Vaughn, S., & Tyler, B.-J. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities, 35*, 386.

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. Unpublished report to the W. T. Grant Foundation, University of Virginia, Charlottesville.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Coffey, J. H., & Horner, R. H. (2012). The sustainability of schoolwide positive behavior interventions and supports. *Exceptional Children, 78*, 407-422.

Cohen, D. & Ball, D. L. (1999). *Instruction, capacity, and improvement.* CPRE Research Report Series RR - 43. Philadelphia: Consortium for Policy Research in Education.

Connor, C. M. (2013). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly, 28*, 342-346. doi: 10.1037/spq0000045

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24*, 1408-1419. doi: 10.1177/0956797612472204

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 by instruction interactions on first graders' literacy development. *Child Development, 80,* 77–100.

Crawford, L., Carpenter, D. M. III., Wilson, M. T., Schmeister, M., & McDonald, M. (2012). Testing the relation between fidelity of implementation and student outcomes in math. *Assessment for Effective Intervention.* Advance online publication. doi: 10.1177/1534508411436111

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical psychology review, 18*(1), 23-45.

Doabler, C.T., Baker, S. K., Kosty, D., B., Clarke, B., Miller, S. J., & Fien, H. (in press). Examining the association between explicit mathematics instruction and student mathematics achievement, *Elementary School Journal.*

Domitrovich, C. E., Gest, S. D., Jones, D., Gill S., & Sanford DeRousie, R. M. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly, 25,* 284-298.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "Implementation research in early childhood education". *Early Childhood Research Quarterly, 25,* 3, 348-357.

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,* 327-350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18,* 237-256.

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

Fishman, J. J., Marx, R., W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and Teacher Education, 19,* 643-658.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R., (2011). *Program evaluation: Alternative approaches and practical guidelines* (fourth edition). Upper Saddle River, NJ: Pearson Education, Inc.

Fixsen, D.L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19,* 531-540. doi: 10.1177/1049731509335549

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231). Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Flay, B., Biglan, A., Boruch, R., Castro, F., Gottfredson, D., Kellam, S., ... Ji, P. (2005). Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination. *Prevention Science, 6,* 3, 151-175.

Fritz, K. (n.d.). Direct Instruction: Technical assistance form. Unpublished instrument. Retrieved from http://www.iup.edu/page.aspx?id=134567

Fuchs, D., McMaster, K., Sáenz, L., Kearns, D., Fuchs, L. Yen, L., … Schatschneider, C. (2010, June). Bringing educational innovation to scale: Top-down, bottom-up, or a third way? Presented at the IES Conference, Washington, DC.

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89,* 253-300Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

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

Good, R., H. III, & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed. rev.). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved from http://dibels.uoregon.edu.

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., & Zins, J. E. (2005). The study of implementation in school-based preventive interventions: Theory, research, and practice. *Promotion of Mental Health and Prevention of Mental and Behavior Disorders (Vol. 3).*Washington DC: U.S. Department of Health and Human Services.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Diffusion of innovations in health service organizations: A systematic literature review. Oxford: Blackwell.

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

Guskey, T. R. (2000). *Evaluating Professional Development* (pp. 67-93). Thousand Oaks, CA: Corwin Press.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Hamre, B. K., Goffin, S. & Kraft-Sayre, M. (2009). *Classroom Assessment Scoring System (CLASS) Implementation guide.* Retrieved from http://www.classobservation.com/wp-content/uploads/2010/06/CLASSImplementationGuide.pdf

Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downter, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly, 25,* 329-347.

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*. Paper presented at the Society for Research in Child Development, Boston, MA.

Hamre, B., Pianta, R., Hatfield, B., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development, 85,* 1257-1274.

Harn, B. A. (2013). QIDR: Examining the Quality of Intervention Delivery and Receipt. Unpublished instrument.

Harn, B. A., Chard, D. J., Biancarosa, G., & Kame'enui, E. J. (2011). Coordinating instructional supports to accelerate at-risk first-grade readers' performance: An essential mechanism for effective RTI. *Elementary School Journal, 112,* 2, 332-355.

Harn, B. & Parisi, D. (2013). The role of fidelity in implementing evidence-based practices in schools. *Savage Controversies*, *6*(2), 2-7.

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,* 181-193.

Harn, B. A., Spear, C. F., Fritz, R., Berg, T. R., & Basaraba, D. (2014, February). *Examining the Relation of Features of Implementation to Student Outcomes.* Conference presentation at the Pacific Coast Research Conference. San Diego, CA.

Heartland Area Education Agency (2008). Implementation integrity direct observation checklist. Unpublished instrument. Retrieved from http://www.lasecfp.org/Intervention-Fidelity-Checklists.html

Howell, K. W., & Nolet, V. (2000). *Curriculum-based evaluation: Teaching and decision making*. Belmont, CA: Wadsworth.

Hox, J. J. (2010). Multilevel analysis: Techniques and Applications (2nd ed.) New York, NY: Routledge.

Implementation check: Reading Mastery (n.d.) Unpublished instrument. Retrieved from http://www.lasecfp.org/Intervention-Fidelity-Checklists.html

Jones, N.D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39,* 112-124. doi: 10.1177/1534508413514103

Kennedy, C. H. (2005). *Single-case designs for educational research.* Boston, MA: Pearson Education.

Klingner, J. K., Ahwee, S., Pilonieta, P., & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. *Exceptional Children, 69*, 411-429.

Klingner, J., Boardman, A., & McMaster, K. (2013). What does it take to scale up and sustain evidence-based practices? *Exceptional Children, 79,* 195-211.

Knoche, L. L., Sheriden, S. M., Edwards, C. P., & Osborn, A. Q. (2010). Implementation of a relationship-based school readiness intervention: A multidimensional approach to fidelity measurement for early childhood. *Early Childhood Research Quarterly, 25*, 299-313.

Kroesbergen, E. H., & Van Luit, J. E. (2003). Mathematics interventions for children with special educational needs. *Remedial & Special Education, 24*(2), 97-114.

La Paro, K. M. , Hamre, B. K., Locasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D., . . . Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development, 20*, 657-692. doi: 10.1080/10409280802541965

Lieber, J., Butera, G., Hanson, M., Palmer, S., Horn, E., Czaja, C., . . . Odom, S. (2009). Factors that influence the implementation of a new preschool curriculum: Implications for professional development. *Early Education and Development, 20*, 456-481.

Luke, Douglas, A. (2004). *Multilevel Modeling*. Thousand Oaks, CA: Sage.

Maas, C. J. & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis, 46,* 427-440.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., . . . Howes, C.. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732-749. doi: 10.1111/j.1467-8624.2008.01154.x

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147-163. doi: 10.1037/1082-989X.9.2.147

Maxwell, K. L., McWilliams, 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.

McGinty, A. S., Justice, L. M., Piasta, S. B., Kaderavek, J., & Fan, X. (2012). Does context matter? Explicit print instruction during reading varies in its influence by child and classroom factors. *Early Childhood Research Quarterly, 27,* 77-89.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315-340.

NICHD ECCRN. (2002). *Classroom Observation System-First Grade*. Charlottesville, VA: University of Virginia.

NICHD ECCRN. (2004). *Classroom Observation System-Fifth Grade*. Charlottesville, VA: University of Virginia.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in k-12 curriculum intervention research. *Review of Educational Research, 78*, 33-84. doi: 10.3102/0034654307313793

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

Odom, S. L., Cox, A. W., & Cook, M. (2013). Implementation science, professional development, and autism spectrum disorders. *Exceptional Children, 79,* 233-251.

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., . . . Marquis, J. (2010). Examining different forms of implementation and in early childhood curriculum research. *Part of special section on Implementation research in early childhood education, 25,* 314-328. doi: 10.1016/j.ecresq.2010.03.001

Odom, S. L., Hanson, M. J., Lieber, J., Diamond, K., Palmer, S., Butera, G., & Horn, E. (2010). Prevention, early childhood intervention, and implementation science. In B. Doll, W. Pfhol, & J. Yoon (Eds.), Handbook of youth prevention science. New York: Routledge.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Harcourt Brace

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109-119. doi: 10.3102/0013189x09332374

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System.* Baltimore: Paul H. Brookes.

Power, T. J., Blom-Hoffman, J., Clarke, A. T., Riley-Tillman, T. C., Kelleher, C., and Manz, P. H. (2005). Reconceptualizing intervention integrity: A partnership-based framework for linking research with practice. *Psychology in the Schools, 42,* 495-507. doi: 10.1002/pits.20087

Raudenbush, S. W. (2009). The Brown legacy and the O'Connor challenge: Transforming schools in the images of children's potential. *Educational Researcher, 38,* 169–180. doi:10.3102/ 0013189X09334840

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

Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org.

Ringwalt, C. L., Ennett, S., Johnson, R., Rohrbach, L. A., Simons-Rudolph, A., Vincus, A., & Thorne, J. (2003). Factors associated with fidelity to substance use prevention curriculum guides in the nation's middle schools. *Health Education & Behavior, 30,* 375-391.

Rosenshine, B. (1997). Advances in research on instruction. In J. Llyod, E.J. Kame'enui, & D. Chard (Eds.) *Issues in educating students with disabilities* (197-221). Mahwah, N.J: Erlbaum.

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

Scientific Software International, Inc. (2012). *HLM 7 student edition.* Retrieved from http://www.ssicentral.com/hlm/student.html

Semmelroth, C. L. & Johnson, E. (2013). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention.* Advance online publication. doi: 10.1177/1534508413511488

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

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children, 31*, 351-380.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 27,* 316-328.

Snijders, T. A. B. (2005). Power and sample size in multilevel modeling. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science.* New York: Wiley.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science, 7*, 43-56. doi: 10.1007/s11121-005-0020-3

Stichter, J. P., Lewis, T. J., Richter, M., Johnson, N. W., and Bradley, L. (2006). Assessing antecedent variables. The effects of instructional variables on student outcomes through in-service and peer coaching professional development models. *Education and Treatment of Children, 29,* 665-592.

Stichter, J. P., Lewis, T. J., Whittaker, T. A., Richter, M., Johnson, N. W., & Trussell, R. P. (2008). Assessing teacher use of opportunities to respond and effective classroom management strategies: Comparisons among high- and low-risk elementary schools. *Journal of Positive Behavior Interventions, 11,* 68-81.

Stith, S., Pruitt, I., Dees, J., Fronce, M., Green, N., Som, A., & Linkh, D. (2006). Implementing community-based prevention programming: A review of the literature. *Journal of Primary Prevention, 27,* 599–617.

Stoolmiller, M., Eddy, J. M., & Reid, J. B. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Counseling and Clinical Psychology, 68,* 296-306.

Sutherland, K. S., Alder, N., & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with EBD. *Journal of Emotional and Behavioral Disorders, 11,* 239-248.

Swanson, H. L., & O'Connor, R. (2009). The role of working memory and fluency practice on the reading comprehension of students who are dysfluent readers. *Journal of Learning Disabilities, 42*, 548-575. doi: 10.1177/0022219409338742

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis. *Review of Educational Research, 68*(3), 277-321.

Tomcho, T. J., & Foels, R. (2012). Meta-analysis of group learning activities: Empirically based teaching recommendations. *Teaching of Psychology, 39,* 159-169. doi: 10.1177/0098628312450414

Vaughn, S., Gersten, R. M., & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children, 67*(1), 99 - 114.

Webster-Stratton, C., Reinke, W. M., Herman, K.C., & Newcomer, L. L. (2011). The Incredible Years Teacher Classroom Management training: The methods and principles that support fidelity of training delivery. *School Psychology Review, 40*, 509-529.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation, 30*(1), 44-61.