EXAMINING MULTIPLE DIMENSIONS OF FIDELITY AND THEIR RELATION TO STUDENT READING OUTCOMES: A RETROSPECTIVE ANALYSIS OF KINDERGARTEN INTERVENTIONS

by

DANIELLE MARIE PARISI

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

June 2009

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

Danielle Parisi

Title:

"Examining Multiple Dimensions of Fidelity and their Relation to Student Reading Outcomes: A Retrospective Analysis of Kindergarten Interventions"

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

Elizabeth Harn, Co-Chairperson, Special Education and Clinical Sciences
Kenneth Merrell, Co-Chairperson, Special Education and Clinical Sciences
David Chard, Member, Special Education and Clinical Sciences
Yvonne Braun, Outside Member, Sociology

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

June 13, 2009

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

An Abstract of the Dissertation of

Danielle Marie Parisi for the degree of Doctor of Philosophy

in the Department of Special Education and Clinical Sciences

to be taken June 2009

Title: EXAMINING MULTIPLE DIMENSIONS OF FIDELITY AND THEIR RELATION TO STUDENT READING OUTCOMES: A RETROSPECTIVE ANALYSIS OF KINDERGARTEN INTERVENTIONS

Approved: ____________________________________
Elizabeth A. Harn, Ph.D., Co-Chair

Approved: ____________________________________
Kenneth W. Merrell, Ph.D., Co-Chair

This dissertation study explored the dimensions of fidelity to aid both researchers and practitioners in their measurement of the construct and use of the data. Understanding the dimensions of fidelity is important for three reasons: (a) limited agreement on a definition, (b) variability in measurement, and (c) inconsistent relations demonstrated between fidelity and outcomes. Leaders in the fields of program evaluation, behavioral health, psychology, and education have begun to promote an expanded definition of fidelity that looks beyond whether surface level components of interventions were delivered to include examination of whether interventions are delivered with quality and whether students are engaged. With this issue in mind, an expanded definition of fidelity was used to explore surface/content dimensions of fidelity or total fidelity, quality/process dimensions of fidelity, and student engagement. Specifically, this study examined how these dimensions relate to each other and how each dimension relates to student literacy outcomes. Multi-process multi-level models were used to study the interrelations among the dimensions of fidelity and the interrelations among the group level fidelity measures and multiple measures of student literacy development.

The results of this study indicated that the construct of fidelity is multidimensional and potentially more complicated than has been discussed in the literature to date. When examining the relations among the dimensions of fidelity, total fidelity and quality were highly related, quality and engagement may be related, and total fidelity and engagement were not related. The relation between total fidelity and student outcomes was in the opposite direction of what was hypothesized—lower total fidelity was related to higher student outcomes. The relation between student engagement and student outcomes was in the hypothesized direction—higher engagement was related to higher student outcomes. The relation between quality of delivery and student outcomes was also in the hypothesized direction with higher quality related to higher student outcomes. The results highlight several issues related to fidelity that need to be considered by both researchers (measuring multiple components, repeated assessment, data analytic methods) and practitioners (how and what to measure, general variability in implementation, use of the data) in the field of education.

CURRICULUM VITAE

NAME OF AUTHOR: Danielle Marie Parisi

PLACE OF BIRTH: Palos Heights, IL

DATE OF BIRTH: January 3, 1982

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
University of Chicago

DEGREES AWARDED:

Doctor of Philosophy in School Psychology, 2009, University of Oregon
Master of Science in Special Education, 2007, University of Oregon
Bachelor of Arts in Psychology, 2003, University of Chicago

AREAS OF SPECIAL INTEREST:

Systems Level Variables Impacting Student Academic Achievement

PROFESSIONAL EXPERIENCE:

Practicum Student, South Lane School District, Cottage Grove, 2007-2008.

Research Assistant and Intervention Coach, Center on Teaching and Learning, University of Oregon, Eugene, 2005-2007.

Practicum Student, Eugene School District 4J, Eugene, 2005-2006.

Practicum Student, Bethel School District, Eugene, 2004-2005; 2006-2007.

Teacher Aide, Quin Early Learning Center, North Palos School District 117, Palos Hills, 2003-2004.

Research Assistant, Goldin Meadow Psychology Laboratory, University of Chicago, Chicago, 2002-2003.

Teacher Assistant/Tutor, Neighborhood Schools Program, University of Chicago, Chicago, 2001-2003.

GRANTS, AWARDS AND HONORS:

Certificate of Recognition, Oregon School Psychologists Association, 2007

Liz Gullion Award, Oregon School Psychologists Association, 2006

PUBLICATIONS:

Harn, B. A., Chard, D. J., Kame'enui, E. J., Allen, M., & Parisi, D. (in press). School-level reading experiences: Examining the role of instructional practices in preventing long-term reading difficulties. *Learning Disabilities Research & Practice.*

Parisi, D. M. & Harn. B. A., (in press). Collaborative measurement of fidelity in schools. *School Psychology Forum: Research in Practice.*

Merrell, K. W., Parisi, D. M., & Whitcomb, S. A. (2007). *Strong Start: A Social and Emotional Learning Curriculum.* Baltimore, MD: Paul H. Brookes Publishing Co.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and thanks to Dr. Beth Harn for her encouragement, support, and guidance throughout the dissertation process and my time at the University of Oregon. I feel privileged to have had the opportunity to work with and learn from her for the past four years. I would also like to thank Dr. Ken Merrell for his support and guidance throughout this project and my time in the program. Thanks also go to Dr. David Chard and Dr. Yvonne Braun for providing feedback and insight as members of my dissertation committee. I would also like to thank Dr. Mike Stoolmiller for providing his expertise, time, effort, and teaching. Thank you also to Dr. Deborah Simmons and Dr. Edward Kame’enui for allowing me to use the Project Optimize data set.

Thank you to my family, especially Mom, Grandpa, Grandma, Michael, and Elise, for being so encouraging and understanding throughout the dissertation process and graduate school.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

Defining Fidelity 2
Expanding the Definition of Fidelity 4
The Importance of Measuring Fidelity in the School Setting 6

II. LITERATURE REVIEW 10

The Role of Instruction in Student Learning and Response to Intervention 10
The Importance of Fidelity for Response to Intervention 14
Defining and Measuring Fidelity 15
Evolution of Fidelity in the Research Setting 15
Expanding the Definition of Fidelity 17
Methods for Assessing Fidelity 19
Indirect Assessment 20
Direct Assessment 21
Relating Fidelity to Outcomes 23
State of the Field in Measuring Fidelity 23
Empirical Examples of Relating Fidelity to Student Outcomes 26
Review of Studies 38
Implications for Research and Practice 39
Refining Interventions 40
Creating Targeted Professional Development Opportunities 42
Helping Schools Document Fidelity 43

III. METHODS 44

Participants 44
Students 44
Interventionists 45
Setting 46
Independent Variable 47
Total Fidelity 47
Quality of Delivery 48
Student Engagement 48
Student Measures 49
Onset Recognition Fluency 49
Phoneme Segmentation Fluency 49
Nonsense Word Fluency 50

Chapter Page

Woodcock Reading Mastery Test-Revised: Word Attack Subtest (Word AT) 51
Woodcock Reading Mastery Test-Revised: Word Identification Subtest (Word ID) 51
Oral Reading Fluency 52
Procedures 52
Interventionist Training 52
Data Collector Training 53
Fidelity Observations 53
Student Assessments 53
Intervention Implementation 54
Measuring Fidelity 54
Measuring Dependent Variables 55
Data Analysis 55

IV. RESULTS 58

Preliminary Analyses and Descriptive Statistics 58
Research Question 1: Relating Dimensions of Fidelity 60
Individual Fidelity Growth Models 60
Engagement 61
Quality 61
Total Infidelity 63
Multi-Process Fidelity Model 63
Research Question 2: Relating Dimensions of Fidelity to Outcomes 65
Fidelity and May Outcome Models 65
Infidelity Predicting May Outcome 66
Engagement Predicting May Outcome 67
Quality Predicting May Outcome 68

V. DISCUSSION 71

Summary of Results 72
Change and Reliability of Dimensions Across Time 72
Research Question 1: Relating Dimensions of Fidelity 72
Research Question 2: Relating Dimensions of Fidelity to Outcomes 73
Limitations 74
Implications for Research 76
Measurement Issues 77
Purpose of Fidelity Collection 78
Relating Fidelity to Outcomes 79

Chapter Page

Implications for School Practice.... 81
Measurement Issues.... 81
Role of Fidelity in School 84
Relating Fidelity to Improved Outcomes 84
Conclusions and Potential Next Steps.... 85

APPENDICES

A. TABLE OF EMPIRICAL EXAMPLES OF RELATING FIDELITY TO STUDENT OUTCOMES 86
B. FIDELITY OBSERVATION FORMS 93
C. RESULTS TABLES 100
D. RESULTS FIGURES 111

BIBLIOGRAPHY 116

LIST OF FIGURES

Figure Page

D1. Trimmed and Transformed Fidelity Constructs 112

D2. Total Infidelity Growth Curves for Each Group 113

D3. Trimmed Quality Growth Curves for Each Group 114

D4. Engagement Growth Curves for Each Group 115

LIST OF TABLES

Table Page

A1. Empirical Examples of Relating Fidelity to Student Outcomes.....87

C1. Descriptive Statistics for Fidelity Measures at the Individual Observation Level 101

C2. Engagement Growth Model Results 102

C3. Quality Growth Model Results 103

C4. Infidelity Growth Model Results 104

C5. Multi-Process, Multi-Level Fidelity Model 105

C6. Between and Within Group Variance Components for May Outcomes 106

C7. Infidelity Predicting Group Level Student Outcome in May 107

C8. Engagement Predicting Group Level Student Outcome in May 108

C9. Quality (Average Level and Slope) Predicting Group Level Student Outcome in May.....109

C10. Quality (Initial Status and Slope) Predicting Group Level Student Outcome in May.....110

CHAPTER I

INTRODUCTION

Learning is an interactive process that is a result of both student characteristics and environmental events. To make decisions that will improve student learning, educators need to consider those variables most pertinent to the educational setting, the curriculum and instruction (Ysseldyke & Christenson, 1988). Traditionally, when student learning is successful, it is assumed that the instruction and curriculum are meeting students' needs. However, when a student does not respond adequately to the schooling experience, an assumption that the student is learning disabled is often made. Current legislation specifically requires a shift away from this process of focusing on within-child disability toward carefully ruling out the contextual variables that may be impacting student learning prior to assuming the student is learning disabled (Vaughn & Fuchs, 2003). The Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004) requires that the quality of the instruction provided, or fidelity, be evaluated to ensure that a student has received a "high quality instructional experience" before considering whether the student has a learning disability. Determining, or quantifying through measurement, the fidelity of these instructional experiences poses a challenge to field for three reasons: (a) there is limited agreement on a definition of fidelity, (b) varying methods are used to measure the construct, and (c) inconsistent relations between fidelity and outcomes have been demonstrated. This section will provide a commonly accepted definition of fidelity, a general overview of how that definition developed based upon work in research settings, and a description of how this understanding of fidelity may impact school-based settings.

Defining Fidelity

The importance of determining the fidelity of interventions in the fields of education, psychology, program evaluation, and behavioral health arose in the research setting and is commonly accepted. The definition most commonly used for fidelity in research studies is the degree to which a treatment condition is implemented as intended (Moncher & Prinz, 1991; Yeaton & Sechrest, 1981). The goal of measuring fidelity is to determine, with a level of confidence, whether the outcomes obtained from a treatment or intervention were in fact related to the intervention and not to other extraneous variables (Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000). As Gresham, Gansle, Noell, Cohen, and Rosenblum (1993) discuss, observing fidelity assists educators in distinguishing between ineffective treatments and effective treatments that may have been implemented with low fidelity. Interventions often are not implemented as designed, and any changes made may have implications for the conclusions that can be drawn from the study. Measuring fidelity in a research study helps researchers to document and address changes to the implementation of an intervention (Lane, Bocian, Macmillan, & Gresham, 2004). It can also help researchers to understand the limits of interventions and their generalizability to other populations and settings (LeLaurin & Wolery, 1992) as documentation that an intervention was implemented as designed aids in establishing a study's external validity as well as in replication efforts (Gresham et al., 2000; Gresham, Gansle, & Noell, 1993; Lane et al., 2004; Moncher & Prinz, 1991).

Though there is agreement on the importance of measuring and documenting fidelity and on a definition, this definition is limited in scope, focusing on whether key pieces of an intervention were implemented and impeding examination of quality of delivery. It is possible that this limited definition of fidelity has led to a lack of agreement amongst researchers and school-based practitioners on how best to measure it as well as on the role and influence of fidelity on student outcomes.

Measurement of fidelity is not consistently reported in the literature. Findings from reviews of studies involving children in the *Journal of Applied Behavior Analysis,* and school-based behavioral interventions across several other journals, indicate that only about one-third of studies reviewed operationally defined the independent variable and a vast majority neither monitored fidelity nor reported fidelity data (Gresham et al., 1993; Gresham, Gansle, Noell, et al., 1993). In addition, a review of learning disability intervention studies revealed that approximately half described fidelity while only about one-fifth measured and reported fidelity data (Gresham et al., 2000). This trend has continued with the majority of intervention studies still not reporting fidelity data. In studies that do measure and report fidelity data, the construct is not consistently defined and measured. Studies use teacher logs, teacher self-report, ratings of permanent products, and direct observations as measures of fidelity. Some researchers are beginning to provide evidence that fidelity is related to student outcomes; however, thus far this relationship has been inconsistently demonstrated.

Expanding the Definition of Fidelity

An expanded definition of fidelity could help both researchers and especially practitioners to better understand how to best measure the construct and use the data to improve student outcomes. Researchers in the fields of education and psychology have begun to broaden the scope of the definition of fidelity by including a focus on quality of delivery as well as student responsiveness to treatment or interventions. Gersten et al. (2005) discuss fidelity in terms of both surface fidelity and quality of delivery while Power et. al (2005) use different terminology, but approach the concepts in the same manner discussing content and process dimensions of fidelity.

Surface or content dimensions of fidelity require an objective look at whether important pieces, established by the researcher/author *a priori*, of the intervention were delivered. These can range from determining any of the following: (a) if central components/features were delivered, (b) if the time allocated was consistent with what was expected, (c) if the intervention was completed (i.e., expected material was covered) and (d) if objectives of the program were adhered to (Gersten et al., 2005; Power et al., 2005).

On the other hand, examining quality of delivery, or process dimensions, requires varying levels of inference. Rather then simply determining if the intervention occurred or a component was delivered, observers attempt to rate how well or to what degree the intervention or component was delivered. Some researchers have stated that quality of delivery may be more directly relevant to outcomes (Gersten et al., 2005), though it will be more subjective and possibly more difficult to capture (Mowbray, Holter, Teague, & Bybee, 2003). Some examples include rating not only how the intervention was delivered, but also qualifying how the student/recipient of the intervention behaves while receiving the intervention. Needless to say, attempting to reliably and consistently capture this type of information across multiple raters will pose challenges for the field, yet it should be considered in relation to outcomes (Power et al., 2005).

The fields of education and psychology are beginning to expand the definition of fidelity to include variables related to whether key components of the intervention are covered, the quality with which the intervention is delivered, and student engagement or receipt of interventions. For the purpose of this study, we define fidelity as *the degree to which central surface level intervention components are implemented with quality such that students are engaged in the intervention.* This definition may assist in our understanding of fidelity in research and educational settings, as the two settings are mutually dependent on one another (Klingner, 2004). Regardless of whether a surface/content or quality/process approach to examining fidelity is taken, this approach needs to be applicable to schools and must inform outcomes and instructional decision-making. The challenge for the field is to determine which dimensions of the construct of fidelity are related to student outcomes and should, for that reason, be systematically measured.

The Importance of Measuring Fidelity in the School Setting

IDEA 2004 has challenged schools by highlighting the necessity of measuring and examining the instructional context when considering eligibility for special education services under the category of learning disabilities. Using a Response to Intervention (RTI) methodology, as delineated in IDEA 2004, involves a prevention-focused system and the implementation of evidence-based interventions to determine student need and eligibility for services. Interventions at all levels of intensity across general and special education need to be implemented with fidelity before educators make high-stakes decisions such as determining whether or not a child is learning disabled (Batsche et al., 2006). Additionally, having procedures in place to measure fidelity on a regular basis helps to ensure that interventions are delivered and instructional support is being provided (LeLaurin & Wolery, 1992). Educators have often skipped this step by assuming that if the student's outcomes were improved, the intervention had been delivered with fidelity; however, the level of behavior change (i.e., improvement in outcomes) may have been even more significant had the treatment been implemented with higher fidelity (Gresham, Gansle, Noell et al., 1993). In schools, we can no longer assume that interventions have been implemented as expected. We must systematically document fidelity to ensure that interventions are implemented with the highest level of quality possible which will lead to important and socially valid levels of behavior change and learning.

Beyond implications for RTI, examining fidelity at the school or systems-level aligns with society's increased emphasis on accountability within No Child Left Behind

(NCLB). Federal mandates such as NCLB and IDEA call for vast changes in curriculum, instruction, and decision-making in schools with a focus on improving student performance (Harn, Chard, Kame`enui, Allen, & Parisi, in press). The field of education as a whole has often been susceptible to fads and frequent adoption of curriculum and reform efforts because they assume the lack of student improvement (if considered) was due to the trend of the day (Vaughn & Dammann, 2001). Yet, most have never examined the quality of implementation of such efforts and may falsely assume that outcomes are related to the latest project when they may be due to the project not being implemented as expected (Borman, Hewes, Overman, & Brown, 2003). Measuring fidelity helps educators distinguish between ineffective treatments and effective treatments implemented with poor fidelity (Gresham, Gansle, Noell, et al., 1993), allowing for those practices that have been implemented with low levels of fidelity to be improved and reconsidered. In addition, our knowledge base of effective interventions, especially in the area of reading, is far more developed than in the recent past; however, typical schools and classrooms are not yet implementing these evidence-based practices at high rates (Denton, Vaughn, & Fletcher, 2003). Monitoring the degree to which school systems and individual teachers implement reform efforts and interventions may assist in bridging the gap between research and practice by identifying areas for professional development (Gersten et al., 2005).

For schools to meet the ever-increasing expectations for all students to be successful across all academic domains (math, science, social studies, etc.), focused and targeted professional development will be essential. As schools continue to strive to meet these growing expectations, procedures and methods for collecting useful fidelity information for teachers will be helpful in this process. On-going progress monitoring of intervention delivery can be helpful in two ways: (a) support can be provided to individual teachers having difficulty implementing new practices through "coaching," and (b) our understanding of the role of variability in implementation over time will be increased. Contextualized feedback provided within a "coaching" approach to collecting fidelity data can allow for timely and individualized support to teachers in a less intimidating manner than other procedures (Chard & Harn, in press) and may be more cost-effective (Moncher & Prinz, 1991). It is expected that modifications will be made to implementation be it purposeful or accidental; identifying areas in need of professional development in a timely basis can improve delivery and related student outcomes (Gresham et al., 2000; Lane et al., 2004; LeLaurin & Wolery, 1992). This on-going support will maximize fidelity leading to improved student outcomes and greater likelihood of sustained use of evidence-based practices (Harn et al., in press). An expanded definition of fidelity using both a surface/content and quality/process approach allows for addressing the components, quality of delivery, and student engagement in relation to student outcomes.

We know that fidelity is important and needs to be measured in schools; however, a persistent challenge in education is bridging the research-to-practice gap and potentially creating a reciprocal relationship in which information gained in the field informs subsequent research (Klingner, 2004). For example, within research applications, all facets of an intervention may be implemented with a high degree of positive outcomes, but are all facets necessary or equally important for those outcomes? Might the intervention be implemented in a manner that is easier for schools yet achieves the same outcomes? Studying the role of surface and quality dimensions of fidelity in effectiveness studies in practical settings to determine which dimensions are important for improving student outcomes can help to enhance intervention effectiveness (O'Donnell, 2008) and to refine the way that fidelity is discussed and measured throughout the field of education.

Researchers have provided much support for the measurement of fidelity in practical settings and have measured fidelity in various ways. In addition, researchers have indicated that systematically studying fidelity in the research setting can help determine the level of implementation that is necessary to achieve an outcome (Halle, 1998) and have argued for the utility of measuring fidelity in practical settings. However, even within the research setting, the possible relation between fidelity and outcomes is unclear. The challenge for the field is to understand the relation between fidelity and student outcomes and to determine which dimensions of fidelity are most relevant to student outcomes. This study looks to fill these gaps in the literature by answering the following questions:

1. What is the relation between dimensions of fidelity (total fidelity, quality of delivery, student engagement)?

2. What is the relation between dimensions of fidelity and student outcomes measured using multiple early literacy measures?

CHAPTER II

LITERATURE REVIEW

The Role of Instruction in Student Learning and Response to Intervention

Learning, though influenced by student characteristics and behaviors, is clearly affected by instruction. Students walk into school with varying skills, strengths, and deficits, but research has identified evidence-based best practices that ensure that all students, even those most at-risk, learn when provided systematic instruction (e.g. Carnine, Silbert, Kame'enui, & Tarver, 2004; Haager, Klingner, & Vaughn, 2007). In a review of prevention and intervention studies in the area of reading, Torgesen (2001) presents two major conclusions: (a) prevention efforts are needed to eliminate reading difficulties and (b) older children need interventions that are "appropriately focused and sufficiently intensive to improve their skills in a short period of time" (p. 199). In offering these conclusions, Torgesen explains that reading difficulties can be both prevented and remediated through the provision of intensive, research-based instruction. Instructional practices and curriculum choice are variables that are under direct control of educators (Howell & Nolet, 2000). Though there is a tendency to focus on student characteristics and look for within-child deficits whenever a problem arises, curricular and instructional variables can be easily altered to ensure that students' needs are met and their overall educational outcomes improved.

IDEA 2004 emphasizes quality instruction when describing an RTI approach for determining whether or not a student qualifies for special education services under the category of learning disability. Instead of a focus on internal student deficits, this new legislation calls for examination of the instructional environment prior to labeling a student as learning disabled. In effect, schools must rule out the possibility that a student's difficulties are related to instruction (Vaughn & Fuchs, 2003). With the passage of IDEA 2004, schools are allowed to use an RTI methodology or a process "that determines if the child responds to scientific, research-based intervention" (Sec 300.309 (b) (1) IDEA 2004). Students cannot be identified as learning disabled if the difficulties are due to a lack of "appropriate instruction."

Though this procedure is only suggested and is not yet required by law, the current reauthorization takes a step in the right direction towards providing prevention-oriented services to all students. The RTI approach will help to solve many problems that are inherent in using an IQ-achievement discrepancy to determine the presence of a learning disability (Fuchs, Mock, Morgan, & Young, 2003). The use of RTI requires educators to "provide early intervention, match instruction to the academic needs of students, and monitor student progress with ongoing data-based decision making" (Vaughn, Linan-Thompson, & Hickman, 2003, p. 392). This approach helps to separate students with true disabilities from students who just need more effective and intensive instruction (Fuchs et al., 2003). The RTI approach requires ongoing progress monitoring through formative methods to ensure that students are being provided instruction that is effective in attaining successful outcomes (Vaughn et al., 2003). The fidelity of the instruction being provided to all students must also be monitored. Clearly, an RTI approach has effective instruction at its core.

Several researchers have illustrated the promise of RTI methodology for providing effective instruction that improves outcomes for even the most at-risk students. The following examples illustrate the use of RTI methodology as well as the power of effective instruction to both prevent and remediate reading problems. Harn, Kame'enui, and Simmons (2007) describe a study of kindergarten interventions that their data show can close the gap between the reading skills of the most at-risk students and typically developing peers. Three interventions were implemented from early November to the middle of May to students scoring below the 20th percentile on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Letter Naming Fluency (LNF) and Initial Sounds Fluency (ISF) measures (Good & Kaminski, 2003). Students in the intervention groups were provided with one of three research-based interventions, two developed by the researchers and one commercially available. One of the researcher-developed interventions had a code or phonemic awareness and alphabetic understanding emphasis while the other intervention had both a code and comprehension emphasis. Students who received the researcher-designed interventions were compared to students who also scored below the 20th percentile on the DIBELS measures and received the commercially available intervention program as well as to average achieving students. Results showed that the typical student in each of the groups—both researcher-developed interventions, the commercially available intervention, and the average achievers—all met DIBELS benchmarks on Phoneme Segmentation Fluency (PSF) and Nonsense Word Fluency

(NWF) at the end of the intervention. These researchers have shown not only that students at-risk of reading difficulty can perform similar to average achieving peers but also, through the examination of two researcher-developed interventions as well as a commercially available program, the key instructional variables that need to be implemented in order to achieve these outcomes.

In another study using RTI methodology, Vaughn et al. (2003) provided supplemental reading instruction to 45 second-grade students identified as at-risk for reading problems. Students were identified as at-risk by teacher nomination and their scores on the screening portion of the Texas Primary Reading Inventory. At-risk students received both core instruction as well as 35 minutes of daily supplemental instruction which included a focus on the five major skills of reading development determined by the National Reading Panel (2000). Specific intervention components included: fluency, phonemic awareness, instructional level reading, word analysis, and writing. Students were assessed on reading skills prior to intervention and then over three 10-week intervals during intervention. Students were exited from intervention if they met criteria established *a priori* during one of the testing sessions. After all 30 weeks, only 25% of the students identified as at-risk at the beginning of the study were still considered at-risk. Twenty-three out of the 24 students who met exit criteria after either 10 or 20 weeks maintained their skills in the general education classroom while 16 of the 24 students continued to make gains. Eight of the same 24 students were not able to make additional gains without continued supplemental support. The authors explain that this study illustrates RTI as a workable option for identifying students with learning disabilities. By establishing criteria for length of and exit from intervention and monitoring student progress throughout intervention, the researchers were able to identify those students who were the most at-risk for reading difficulties and who would continue to need support. While these and other studies demonstrate the potential and power of RTI, a lingering concern is how to document the quality, or fidelity, of the instruction provided.

The Importance of Fidelity for Response to Intervention

RTI requires the implementation of research-based interventions to identify students as needing special education services under the category of learning disabilities. Vaughn and Fuchs (2003) explain that RTI allows contextual (i.e. instructional) variables to be eliminated as the explanation for any student's academic difficulties. They assert that "the failure to respond verifies that the deficit resides in the individual, not the instructional program" (pp. 142). For this to be true, researchers and practitioners must ensure that interventions are implemented as planned or with fidelity. Stating that an intervention will happen is not the same as ensuring that it was done well or as specified (Gresham, 1989).

Leaders at the forefront of RTI research have emphasized the role of fidelity in the RTI process. In a report prepared by the National Joint Committee on Learning Disabilities (NJCLD; 2005), key pieces of data necessary in an RTI model are listed. They include documentation of: (a) research-based instruction in general education, (b) intensive implementation of interventions matched to individual student difficulties, (c) collaboration between school staff, (d) monitoring of student progress, (e) parent involvement, and (f) compliance with timelines described in the federal regulations. Fidelity is addressed in their final recommendation: "Systematic assessment and documentation that the interventions used were implemented with fidelity" (p. 2). This group as well as the National Association of the State Directors of Special Education (NASDSE; Batsche et al., 2006) delineate intervention fidelity not only as important but also as a major challenge to RTI implementation. The field of education must figure out both who will measure the construct and how it should be measured to ensure that the most useful information possible is collected. Determining the fidelity of interventions is a challenge for the field for poses a challenge to field for three reasons: (a) there is limited agreement on a definition of fidelity, (b) varying methods are used to measure the construct, and (c) inconsistent relations between fidelity and outcomes have been demonstrated.

Defining and Measuring Fidelity

Evolution of Fidelity in the Research Setting

Many terms have been used to discuss fidelity in the literature including treatment integrity, fidelity of implementation, treatment fidelity, and implementation of the independent variable. For the purposes of this study, it will be discussed as fidelity. Fidelity is measured in research settings for a variety of reasons. At a basic level, fidelity is measured to ensure that interventions were implemented (LeLaurin & Wolery, 1992; Orwin, 2000). In addition, documenting and measuring fidelity aids in demonstrating internal, external, and statistical conclusion validity as well as increased statistical power and effect sizes. Orwin (2000) explains that measuring fidelity allows researchers to determine whether the study was a "good test" of how an intervention should work. When conducting research, the purpose is to document that changes in the dependent variable are due to manipulation of the independent variable or intervention, in other words that there is a functional relation between the independent variable and the dependent variable (Peterson, Homer, & Wonderlich, 1982). To accomplish this, researchers must measure the independent variable to demonstrate that they have control over it (Peterson et al., 1982) and to ensure that the treatment is not being implemented in the control group (Mowbray et al., 2003). Peterson et al. (1982) and Gresham et al. (1993) discuss the "curious double standard" in research studies where dependent variables are systematically and precisely assessed while assessment of the independent variables is ignored. They caution that observation of only the dependent variable does not allow a researcher to account for all of the variability in the dependent variable; assuming that a stable dependent variable indicates stable implementation of the independent variable is not always accurate. Different dosages of an intervention may be required to maintain the same response from a student over time or from different students at the same time (Peterson et al., 1982). Furthermore, because higher internal validity is correlated with higher effect sizes, documentation of fidelity leads to improved internal validity and thus increased effect sizes (Bellg et al., 2004).

When a high level of fidelity is documented, researchers can be more confident in the conclusions that they draw as it removes the possibility that an intervention could have been more effective if it had been implemented with higher fidelity (Yeaton &

Sechrest, 1981). In addition, monitoring fidelity helps to reduce variability in the independent variable which improves statistical power (Bellg et al., 2004; Moncher & Prinz, 1991; Mowbray et al., 2003). Documentation of fidelity also helps to improve external validity by helping researchers to understand the generalizability and limits of interventions (LeLaurin & Wolery, 1992). By documenting how an intervention was implemented, researchers are better prepared to replicate and generalize their findings to applied settings (Gresham et al., 1993; Moncher & Prinz, 1991).

By documenting and measuring fidelity, researchers can distinguish between an ineffective intervention and an intervention that could have been effective but was implemented poorly (Gresham et al., 1993). In the research setting as well as for practical applications of interventions, it is useful to document fidelity or lack thereof as it is quite common for implementers to deviate from prescribed delivery (Gresham, Gansle, Noell et al., 1993; Lane et al., 2004; Mowbray et al., 2003). By documenting such changes, researchers can correct problems early before any possible negative effects occur and rule out poor implementation as a reason for any negative findings or outcomes (LeLaurin & Wolery, 1992; Orwin, 2000; Peterson et al., 1982). Additionally, field-based modifications, while impacting fidelity, may inform the field about better methods of implementation.

Expanding the Definition of Fidelity

As discussed above, researchers began using the concept of fidelity to measure the degree to which a treatment is implemented as intended (Moncher & Prinz, 1991; Yeaton & Sechrest, 1981). In the field of education, fidelity is most often described and measured as the accuracy and consistency with which an intervention is delivered (Lane et al., 2004). This requires an objective look at whether or not the key components of an intervention were implemented. However, the fields of education, psychology, behavioral health, and program evaluation have recently expanded the way that fidelity is discussed in the literature. This may help schools better conceptualize fidelity which, in turn, will inform data collection as part of RTI implementation. An expanded definition of fidelity can also help researchers determine how to better measure the construct within research studies and to understand the relationship between fidelity and outcomes. The definition has been broadened by several authors to include both the quality of delivery of an intervention and student engagement.

In education, Gersten et al. (2005) and Power et al. (2005) have discussed fidelity in terms of surface/content dimensions and quality/process dimensions. The surface/content dimensions include an objective look at whether or not key components of an intervention, as determined by the researcher or author of the program *a priori,* were delivered. To measure these aspects of fidelity, delivery of key features of the intervention, time allocation, exposure to the specified material, and adherence to the objectives of the program are examined (Gersten et al., 2005; Power et al., 2005). The quality/process dimensions include a more subjective look at the quality with which interventions are delivered as well as at student engagement during intervention delivery.

From a program evaluation perspective Mowbray et al. (2001) provide a definition of fidelity that includes a look at intervention delivery related to structure—following a prescribed framework for service delivery—and delivery related to process—

the way that the services are delivered. In a discussion of measuring fidelity within program evaluations of substance abuse programs, Orwin (2000) talks about fidelity in terms of a hierarchy of adherence, participation, and general fidelity. Adherence includes participant attendance and program completion, participation involves a participant being engaged in the intervention and not just attending, and general fidelity is the traditional description of fidelity or "adherence of actual treatment delivery to the protocol originally developed" (Orwin, 2000, p. 310). The adherence and participation components of this discussion of fidelity also include an overt focus on whether participants or, in the case of education, students are receiving intervention content and are engaged in interventions. In a discussion of the concept and measurement of fidelity in health behavior intervention research, Bellg et al. (2004) explain that fidelity has been expanded to encompass treatment receipt and treatment enactment or whether participants are engaged during the delivery of interventions as well as whether they generalize their skills outside of the intervention setting. The fields of education, psychology, behavioral health, and program evaluation are calling for an expansion of the definition of fidelity to include examination of variables related to quality of delivery as well as to participant or student engagement. Though all of the definitions and discussions referred to above are not from the field of education, they clearly apply.

Methods for Assessing Fidelity

Researchers have developed several methods for assessing fidelity. They vary along a continuum of complexity in terms of the ease with which the data is collected and the variables that are examined. Often, the more simple the data collection method, the larger the inference required to interpret the data. When high levels of inference are required, it is more difficult to be certain that the results give a full and accurate picture of intervention implementation. Gresham (1989) provides a review of several methods for assessing fidelity that can be categorized as either indirect or direct assessments of fidelity.

Indirect Assessment

At the more simplistic end of the continuum of complexity, some researchers address fidelity by providing a manualized treatment and script to interventionists. Though this is a very easy way to help promote fidelity, using it as a measure of fidelity requires the assumption that the interventionists adhere to the protocols and scripts. Other researchers have also used permanent products as a measure of fidelity. When using this method, researchers collect teacher products, including attendance forms, or student products, including worksheets, as evidence that an intervention was implemented as planned. Again, this method does not give direct evidence that all components of an intervention were implemented as planned; the researcher is left to infer that they must have been. Gresham explains that some researchers choose to interview teachers after the intervention has taken place or have teachers complete self-report or self-monitoring forms. These methods are problematic because teachers may be inclined to fill out the forms or answer questions in a socially desirable way. In addition, when using self-monitoring techniques, teachers are required to expend their time and energy on monitoring their teaching rather than just teaching.

Direct Assessment

A very basic direct assessment method involves having observers directly observe the implementation of an intervention and fill out a behavior rating scale at the end of the observation period. Gresham cautions that using this method allows for an overall rating of the entire intervention and not a systematic look at each component of an intervention. The most complex and least inferential method for measuring fidelity is direct observations conducted in real time. Direct observations in real time allow for an unbiased observer to systematically observe implementation of the overall intervention as well as its key components on a consistent basis to ensure that the intervention is being implemented as planned.

There are multiple ways to conduct direct observations. At the simplest level, a fidelity protocol including operational definitions of the key components of an intervention should be created. Within this fidelity protocol, overall session fidelity, or the percentage of key components implemented in a session, can be gathered and component fidelity, or the percentage of implementation of a component over multiple sessions, can be documented (Gresham et al., 2000; Lane et al., 2004). This type of direct observation is based on a limited definition of fidelity that primarily looks at surface or content dimensions of fidelity. When expanding the definition to include quality or process dimensions, an observation can include not only a checklist of whether key components were implemented but also a rating of how well the intervention was delivered. Observations can also take into account student response by including observations or overall ratings of student engagement and/or accuracy. Issues of reliability come up in any data collection system because error is always a possibility even in the more simplistic indirect methods of assessment. However, in direct observations, reliability is especially important because the teacher and/or student behavior cannot be repeated if there is a question about accuracy in the observation. There are several factors that may affect data collection using direct observations. They include: (a) reactivity, the presence of an observer can affect an interventionist's behavior; (b) observer drift, observers may veer from the observational protocol as time goes by; (c) complexity, more complex systems are more prone to error; and (d) expectancy, observers may be searching for specific implementation behaviors (Alberto & Troutman, 2003). Therefore, it is imperative that reliability data be collected, especially when conducting direct observations. This data should consider not only inter-observer agreement but also variability in implementation across time (Stoolmiller, Eddy, & Reid, 2000).

Wickstrom, Jones, LaFleur and Witt (1998) conducted a study to assess the effects of behavioral consultation on teachers' fidelity and the relationships between problem severity, treatment acceptability, and degrees of collaboration and teacher fidelity. This study highlights the continuum of complexity and the varying information that is gained using different methods for assessing fidelity. Fidelity was assessed in three ways. The first was scores on the Baseline and Intervention Record Form (BIRF) that teachers were to use to monitor students' behavior. This was considered a measure of fidelity because a goal of the consultation provided was to get teachers to collect data on the behavior of their students. The second measure of fidelity was stimulus, or permanent product, use. An observer noted whether the required intervention stimulus was near the student's desk during two observations. The third measure of fidelity was treatment use which was measured using direct observations of the percentage of target behaviors or alternative responses that were followed by a planned consequence. Scores on the BIRF were the most indirect measures of fidelity as it was a teacher permanent product. The mean fidelity score when using the BIRF was 54%. Stimulus product use as a measure of fidelity required minimal observation and resulted in a mean score of 62%. Using direct observations of treatment use as the measure of fidelity resulted in a score of 4%. The authors point out that estimates of fidelity decreased as the level of methodological rigor increased from indirect (teacher permanent product) to direct measures (direct observation) of fidelity.

Relating Fidelity to Outcomes

Several leaders in the fields of psychology, program evaluation, and education have highlighted the importance of relating fidelity to outcomes (Gersten et al., 2005; Gresham et al., 1993; Perepletchikova & Kazdin, 2005). Before the field of education advocates the consistent measurement of fidelity in schools, we should understand its impact on outcomes as the cost of collecting fidelity data is high (Zvoch, Letourneau, & Parker, 2007). Though there has been a call to relate fidelity to outcomes, most researchers have historically not measured or reported fidelity data.

State of the Field in Measuring Fidelity

In 1982, Peterson et al. reviewed articles in the *Journal of Applied Behavior Analysis* published from 1968 to 1980 to determine whether or not researchers were operationally defining the independent variable and assessing fidelity. The authors examined whether articles published each year provided operational definitions of the independent variable by coding (a) yes, an operational definition was included, (b) no, one was not included and not needed, and (c) no, one was not included but was needed. They found that a majority of studies (~80%) did operationally define the independent variable when necessary; however, in each year, approximately 10 to 50% of articles did not include operational definitions when needed. Of the studies that did present operational definitions of the independent variable, an average of only 16% also measured fidelity.

Building from Peterson et al.'s (1982) study, Gresham et al. (1993) reviewed studies involving children as subjects from the *Journal of Applied Behavioral Analysis* published between 1980 and 1990 to examine how independent variables were described and whether fidelity was measured. They found that out of 158 studies, 34.2% provided an operational definition of the independent variable or intervention, 15.8% systematically measured and reported levels of fidelity, and 8.8% stated that fidelity was monitored but did not provide data. Simultaneously, Gresham, Gansle, Noell, et al. (1993) reviewed school-based behavioral interventions published in seven journals from 1980 to 1990 to again examine how fidelity was treated. Of 181 studies, 35% provided an operational definition of the independent variable, 14.9% measured and reported levels of fidelity, and 9.9% measured fidelity but did not provide any data. Gresham, Gansle, Noell, et al. also documented a significant correlation between percent of fidelity and effect size ($r = .51, p < .05$) and between percent of fidelity and percent of non-overlapping data points in single subject studies ($r = .58, p < .05$). As percent of non-overlapping data points is a measure of effect size in single subject studies, this result provides evidence that higher fidelity is associated with higher effect size. Therefore, in the 1980's and 1990's fidelity was not often considered. This data indicates that only about a third of studies reviewed were operationally defining the independent variable and a vast majority neither monitored fidelity nor reported fidelity data.

More recently, Gresham et al. (2000) performed a review of articles involving interventions in three major learning disabilities journals from January of 1995 to August of 1999 to also explore whether or not fidelity was regularly assessed. These journals included the *Journal of Learning Disabilities, Learning Disability Quarterly,* and *Learning Disabilities Research & Practice.* The authors found that only 18.5% of intervention articles in these three journals measured and reported data on fidelity. Approximately half of the studies reviewed mentioned fidelity but did not provide any numerical data, and over 30% of the articles did not mention fidelity at all.

From the program evaluation literature, Zvoch et al. (2007) reviewed multisite evaluations published in *New Directions for (Program) Evaluation* and found that eight out of nine of the studies collected data related to monitoring fidelity. However, the authors reported that the method for collecting the data and whether the data was used to evaluate the impact of the program was unclear.

Though fidelity is a "hot topic" currently, this has not always been the case. In major psychology and education journals over the passed three decades, fidelity has often been ignored. Therefore, it is difficult to determine whether or not fidelity is directly linked to student outcomes. Furthermore, the above studies do not explicitly explain how fidelity was defined and measured as Zvoch et al. (2007) highlights. The field of education is calling for direct and systematic measurement of intervention implementation to determine fidelity in the school setting; however, it is difficult for schools to make progress in this area if the field cannot agree upon the definition of fidelity and methodology to assess fidelity or find consistent linkages to student outcomes.

Empirical Examples of Relating Fidelity to Student Outcomes

Certainly, there have been studies that have systematically documented fidelity and related this data to student outcomes. However, these studies are few and because of differing definitions of fidelity and differing methodology for measuring the construct, the findings have not been consistent (Perepletchikova & Kazdin, 2005). Some researchers measure only surface/content dimensions of fidelity while some measure both surface/content and quality/process dimensions, and both are measured in varying ways. Following is a review of studies that have systematically measured fidelity and related fidelity scores to outcomes. In the studies, both surface/content and quality/process dimensions have been measured in numerous ways on the continuum of complexity. A summary of the studies, their methods for measuring fidelity, and their results can be found in Table A1 in Appendix A.

Witt, Noell, LaFleur, and Mortenson (1997) conducted a single subject study that examined the use of performance feedback to improve the fidelity with which four general education teachers implemented an academic intervention. To assess fidelity, permanent products were collected and fidelity was calculated as the percentage of correct permanent products received divided by the total number of treatment steps for the day. This method is on the simple end of the continuum of complexity and measures surface dimensions of fidelity. Though the focus of this study was teacher behavior, the authors found that their intervention improved students' academic performance and that higher levels of fidelity resulted in an increase in academic performance for three out of the four students.

Persampieri, Gortmaker, Daly, Sheridan, and McCurdy (2006) conducted two single subject studies of the effects of parent-delivered reading interventions on student outcomes. Within this study, the relationship between fidelity and student outcomes was also examined. To measure fidelity in the first study, sessions were recorded on an audiotape and a researcher listened to 40% of the sessions. The researcher calculated the number of steps completed and divided that by the total number of steps on the intervention protocol. A sticker reward chart was also used as a measure of how often the intervention was implemented. In the second study, parent report was used as the measure of fidelity. Parents were given a fifteen-step protocol and asked to record each step implemented. Parent-lead sessions were audiotaped and reviewed by a researcher. For three of the five subjects across the two studies, correct words read per minute, the outcome measure, decreased during weeks when fidelity was low. All of the methods for assessing fidelity employed in this study focused on the surface dimensions. The sticker chart is considered a permanent product, and it and the self-report measure are indirect methods on the simpler end of the continuum of complexity. Assessing an audiotape is more complex and direct as it involves listening to an entire lesson, though not in real time.

Van Otterloo, van der Leij, and Veldkamp (2006) examined the effects of a home-based phonological awareness intervention on child outcomes and also looked at how fidelity contributed to early reading skills at the end of kindergarten. The researchers measured what they termed "quality" and "quantity of implementation" which is consistent with quality and surface dimensions of fidelity. To observe quality of implementation, a researcher videotaped one tutoring session and analyzed the session using an observation composed of five 5-point Likert scales that measured child persistence, enthusiasm, and responsiveness to parent, parent instruction adapted to the child, and parent supportive presence. This is on the more complex end of the continuum for measuring fidelity as it involves observation of a videotaped lesson. Quantity of implementation was measured using daily log forms on which parents checked the components of the lesson that were completed. This is on the simpler end of the continuum. When analyzing the contribution of fidelity to child outcomes, the five Likert scales loaded on one factor so one quality of administration variable was created. Regression analyses showed that quantity and quality of administration together accounted for 43% of the variance in early reading skills at post-test while the contribution of the quantity measure was larger than that of the quality measure. When no other variables were controlled for, quantity of administration accounted for 36% of the variance. After controlling for the child's receptive vocabulary and the education level of the mother, quantity accounted for more than 30% of the variance. However, after controlling for pre-test early reading skills, quantity accounted for 12% of the variance in the dependent measure. Quality of administration accounted for 10% of the variance when quantity of administration was controlled. When quantity and education level of the mother or pre-test scores were accounted for, quality did not account for any of the variance. The authors point out that their quality measure was more a measure of climate and the interactions between the parent and child rather than just quality of implementation. In addition, they caution that fidelity was only measured once during the study and may not be representative of implementation quality across the entire intervention. They also assert that quantity of implementation was very easy to assess in comparison to quality of administration making it more cost-effective.

Al Otaiba and Fuchs (2006) conducted a study to determine student characteristics that predict responsiveness and nonresponsiveness to early literacy interventions. Within this study, they also looked at the fidelity of effective, research-based early literacy interventions, Ladders to Literacy and Peer Assisted Learning Strategies (PALS), and the relationship of fidelity to student reading outcomes. For K-PALS and 1st grade PALS, fidelity was evaluated five times across kindergarten and first grade. Researchers observed three student pairs that were randomly chosen using a checklist that scored behavior as demonstrated, not demonstrated, or not applicable. An overall classroom score was created by combining the teacher and average student scores from the observation; each student in the study was observed once. For the Ladders to Literacy intervention, teacher calendars were used to determine the number of activities conducted. Both of these measures of fidelity focus on surface level variables. To measure quality of delivery for the Ladders to Literacy intervention, teachers were observed and given a weekly global 1 (poor) to 3 (excellent) rating addressing lesson clarity, how well the teacher's instruction fit the intent of the lesson, and the degree to which all students were engaged. Students were determined to be nonresponsive, sometimes responsive, or always responsive based on their performance across a range of literacy, language, and behavior measures. ANOVAs were conducted to determine the relationship between student responsiveness to intervention and fidelity. Statistically significant differences in the fidelity of Ladders and not PALS were found in relation to student responsiveness status (nonresponsive vs. sometimes responsive vs. always responsive). Post hoc pairwise comparisons using the Tukey HSD method showed that nonresponsive students were in classrooms where K Ladders activities were implemented with lower quality. For example, the mean fidelity score for the eight classrooms in which nonresponsive students were members was 2.10 while the mean fidelity score for the 12 classrooms in which sometimes responsive students were members was 2.39, and the mean fidelity score for the 17 classrooms in which always responsive students were members was 2.39. Nonresponsive students were in classrooms with lower fidelity for first grade PALS in the fall than sometimes and always responsive students.

Studies have also been conducted to assess the impact of fidelity of classroom-, school-, or district-level interventions on student outcomes. Gettinger and Stoiber (2006) conducted a study of the effects of a functional assessment and positive behavior support program on classroom behavior. School-based teams in pre-kindergarten through first grade classrooms implemented FACET, a functional assessment and positive behavior support program; one to two children in each classroom were nominated to participate. Behavioral outcomes of interest were social cooperation, engagement and learning behavior, aggression, distractibility, noncompliance, negative affect, and specific target behaviors unique to each student. Record forms were used by implementers as a self-assessment of fidelity, and the same forms were coded by observers. Each step of the FACET problem-solving program was broken down into 5 to 8 activities that were coded as 0 (not completed), 1 (completed, with minimum specificity), 2 (completed, with sufficient specificity). This can be considered a measure of surface dimensions of fidelity. Correlations between fidelity of each component of FACET and improvement in student behavior ranged from .47 to .77. The correlation between fidelity and grade level was -.46. The program was implemented with higher fidelity for younger children, and children in younger grades also made greater gains in positive behavior. Therefore, the authors concluded that children in grades where fidelity was higher made greater gains in positive behavior.

Telzrow, McNamara, and Hollinger (2000) examined the fidelity of problem-solving implementation using the Intervention Based Assessment (IBA) process by 227 multidisciplinary teams in Ohio and the relationship between fidelity and student outcomes. Surface level dimensions of fidelity were measured using two work products. The first was a problem-solving worksheet that listed all of the problem-solving components; the second was an evaluation team report form. A Likert scale and scoring rubric were used to evaluate the work products, focusing on implementation of the problem-solving components and student outcomes. This can be considered an indirect measure of quality of delivery. Fidelity ratings for six of the eight problem-solving components were significantly but modestly correlated with ratings of student outcomes. The two components with the lowest fidelity ratings were not significantly correlated with student outcomes. A stepwise multiple regression analysis showed that two problem-solving components were significant predictors of student outcomes and accounted for 8% of the variance. The authors caution that, overall, levels of fidelity were moderate which may limit the conclusions that can be drawn. They also explain that years of participation in the IBA process project was not related to fidelity of the program.

Kovaleski, Gickling, Morrow, and Swank (1999) evaluated the effects of high vs. low implementation of the Instructional Support Team (IST) process implemented statewide in Pennsylvania on academic learning time. In evaluating the IST process, the authors hypothesized that improvement in students' time-on-task, task comprehension, and task completion would depend on the school's level of implementation of critical program features. Level of implementation or fidelity data were taken from a validation process that was managed by the state. Data were collected at the end of the schools' second year of implementation. For schools in Phase I of IST implementation, a three-person team from a different part of the state filled out a 103-item checklist that required them to indicate the number of program components in place. This is a measure of surface dimensions of fidelity. For schools in Phase 2 of IST implementation, a tool that had seven broad areas of implementation rated on a 4-point scale was used: 0 (feature not in place), 1 (basic feature in place), 2 (feature in place at effective level), and 3 (feature in place at model level). This is also a measure of surface fidelity. Schools with the top 30% of scores were considered to be high implementation schools while the schools with bottom 30% of scores were considered to be low implementation schools. High implementation schools had higher gains in task comprehension scores than low implementation and non-IST schools while low implementation and non-IST schools did not significantly differ. For task completion, there were no differences between groups from pretest to posttest; however, from posttest to follow-up, high implementation schools showed an increase in task completion, while low and non-IST schools showed a decline. Finally, when examining time on-task, groups did not show significant differences except that low implementation groups had lower scores on time on-task than non IST-schools. From posttest to follow-up, high implementation schools showed more gains on time on-task than low or non-IST groups. The authors point out that "half-hearted" implementation of the IST process was no better than not implementing at all, and over time, student in high implementation schools were beginning to look like their average achieving peers. Further, they explain that they did not examine specific components of the process that were in place but instead used an overall implementation score.

In a much more complex study from a program evaluation perspective, Zvoch et al. (2007) conducted a multisite evaluation of an early childhood literacy problem. Specifically, they were exploring the relationship between fidelity and student literacy outcomes. The Voyager Universal Literacy program was delivered to 1,229 kindergarten students across 49 classrooms in 21 schools in the Southwest for an entire school year.

Forty-nine kindergarten teachers implemented the intervention. Classrooms varied in size and schools were on either a 9 month or year round schedule. To measure fidelity, school district personnel using background knowledge, program manuals, and expert consensus dialogue created a 6-item checklist. Graduate students and retired educators conducted three observations during 2-week windows in October, January, and April. Scores for each teacher were averaged across the three observations to create an overall score. This method assesses surface dimensions and does not get at the quality dimensions of an expanded definition of fidelity.

In their analysis, contextual data on students, teachers, and classrooms were included in multilevel models. Three-level longitudinal growth models were used where observations were nested within students and students were nested within treatment sites. By employing a multilevel analysis, the authors were able to estimate student growth trajectories and directly model the fidelity data while examining whether site or provider characteristics were associated with fidelity. Student outcomes were assessed using the Dynamic Indicators of Basic Early Literacy Skills Initial Sounds Fluency, Letter Naming Fluency, Phoneme Segmentation Fluency, and Nonsense Word Fluency measures. Students' scores on all four of the measures were added together to create a composite score. The authors found that provider characteristics and class size were not significantly related to fidelity. Site-to-site differences accounted for 43% of the variability in students' literacy growth, and student characteristics accounted for 9% of the variance in students' literacy growth and 3% of the variance in site's initial scores and growth. The only variable that predicted differences in student outcomes was the school schedule, either 9-month or year round. Most important for the purposes of the current study, fidelity scores were not related to site-based student growth rates. The authors examined the relationship between fidelity and growth rates more closely by computing conditional growth rates for each site and plotting the bivariate relationship between site growth rates and fidelity scores. They found that the relationship between the two variables was generally positive and that three sites had extreme scores of low implementation and high growth in student outcomes that greatly impacted the relationship. When completing the analysis after removing these three sites, a significant association between fidelity and student outcomes emerged. The authors explain that there were high levels of variability in the student literacy growth rates of sites with high implementation scores. They also point out limitations in that the fidelity observations did not allow for a "fine-grained distinction of the degree to which an observed component was implemented" (p. 145) and that reliability checks on the fidelity observations were not conducted. They also explain that the sample size did not allow for analysis of the relationship between fidelity and outcomes at the classroom rather than the school level. Based on the results of this analysis, the authors discuss the fact that different contextual factors may make deviations and modifications of a program's protocol more effective than strict adherence to it.

In an early Project Follow Through study, Gersten, Carnine, and Williams (1982) published a description of observation tool called the Direct Instruction Supervision Code or DISC and how it was developed, its reliability and concurrent validity, and patterns of teacher and paraprofessional skill development. The tool was created to observe implementation of direct instruction as variability in results from the National Follow Through Study were thought to be “due in part to fluctuations in the extent to which a model was implemented at a given site” (p. 67). The DISC records rates and frequencies of the following seven behaviors hypothesized to be related to student outcomes: (a) accuracy of formats, (b) use of hand signals, (c) use of corrections, (d) pacing of lessons, (e) student accuracy rate, (f) reinforcement, and (g) time allocation. The authors explain that format accuracy applies to any teaching model as “the precision with which teachers follow the curriculum” (p. 69). To establish concurrent validity of the DISC, teachers were also given global ratings on a one to four scale. To determine the utility of using the DISC versus a general teacher interview, teachers were administered the Levels-of-Use interview. Teachers and paraprofessional aides who were participants in Project Follow Through were observed once in November, once in either late December or January, and two to four times in May while teaching their Direct Instruction lessons in reading or language. In terms of temporal stability, results showed that teachers and aides must be observed at least four and possibly more times in one week in order to obtain a stable estimate of performance. To obtain concurrent validity scores, performance scores on each variable of the DISC in the winter were correlated with the global rating. The median correlation was .45. To determine the construct validity of the DISC, two procedures were used: (a) the three highest-ranked and four lowest-ranked teachers were contrasted on spring observational scores and (b) the relationship between scores on the DISC and scores on the Levels-of-Use interview were examined. Using the contrasted groups procedure, results showed that high ranked teachers scored higher on pacing, student accuracy, and use of corrections. On the other hand, there were no differences between high ranked and low ranked teachers on format accuracy, and the authors assert that this is "probably because all teachers had mastered this most basic skill" (p. 73). The lowest ranked teachers had higher scores on the use of hand signals. Based on these results, total DISC scores were computed by averaging a teacher's scores on the three variables that discriminated high from low teachers--pacing, student accuracy, and use of corrections. When examining the relationship between the Levels-of-Use interview and DISC scores, the researchers did not use any formal correlations because of the restricted range of scores on the interview tool. They found, informally, that scores did not correlate well with DISC scores and highlighted several problems with using the Levels-of-Use interview. They explain that it was not sensitive to differences in implementation and that teachers are able to answer questions in a way that would indicate implementation with high fidelity even if they had never implemented the program.

The authors also explored the relationship between the DISC system and student academic performance. The authors used the Total Reading subtest of the Comprehensive Test of Basic Skills (CTBS) as a measure of student outcomes. They found that the classes of the two teachers with the highest DISC scores had CTBS scores above the national median (the 52nd and 59th percentiles) while the classes of the two teachers with the lowest DISC scores had CTBS scores that were low (the 27th and 22nd percentiles). This provides evidence that teachers who implemented Direct Instruction methods with higher fidelity had students with higher outcomes. When looking at patterns of implementers' performance on the DISC, the authors found that paraprofessionals' mean was often a bit lower than teachers'. However, after two months of training, almost all implementers mastered use of formats and signals. When reflecting on the DISC system, the authors explain that it is helpful as a measure of implementation, which is necessary in the summative evaluation of a model. They also explain that a tool that uses direct observation, a direct and more complex methodology, is not confounded by subjectivity on the part of raters or the verbal ability of teachers when they are interviewed. Data from this measurement tool help to target areas needing further training and professional development.

Review of Studies

In the above studies, fidelity was measured in numerous ways and was not consistently defined across studies. All nine studies measured surface dimensions of fidelity while only four of the nine also documented quality dimensions of fidelity. Most studies found a positive relationship between fidelity and outcomes; however, the strengths of these relationships and the methodological rigor with which they were documented varied. The authors of the studies offer several reasons to expect inconsistent findings and issues to consider when relating fidelity to student outcomes. Van Otterloo et al. (2006) and Gersten et al. (1982) caution that assessment of fidelity at one point during an intervention is probably not a reliable measurement of implementation across an entire intervention. They also explain that implementers of interventions may react to the presence of observers and change the way they are implementing the intervention and that teachers may respond in a socially desirable manner when using an indirect method such as teacher interviews to assess fidelity. These issues point to the need to measure fidelity on a consistent basis over the course of an intervention in order to get a reliable estimate of implementation. Kovaleski et al. (1999) caution that measuring overall fidelity does not capture which specific components are in place in an intervention which impacts the conclusions that can be drawn. Zvoch et al. (2007) also point out the need to consider the reliability of the fidelity data collected as well as the level of analysis when interpreting data, and the impact that contextual factors can have on the level of fidelity that is necessary for student success.

Implications for Research and Practice

Researchers in several fields are beginning to explore the relationship between fidelity and student outcomes. However, their definitions for fidelity and methodology for assessing the construct vary. Expanding the definition of fidelity and systematically exploring the relationship between fidelity and student outcomes using sound methodology that requires low levels of inference and is representative of day-to-day implementation of an intervention is imperative to the field of education. Doing so will benefit both research and practice by helping bridge the research to practice gap. In education today through legislation such as No Child Left Behind and IDEA 2004, emphasis has been placed on the use of evidence-based practices. Evidence-based interventions and instructional practices exist but are not often implemented in the classroom (Denton et al., 2003; Vaughn & Damman, 2001). Collection of fidelity data in both research and practical settings can help to bridge this gap by providing support to practitioners through professional development and by helping researchers gain information about the implementation of interventions in real school settings. To bridge this research to practice gap, Denton et al. explain that the following things are needed: "(a) the provision of better linkages between researchers and teachers, (b) support of educational research and development that yields knowledge that is practical and applicable in classrooms, and (c) the provision of clear documentation of practices that are research-based and opportunities for teachers to access this information" (p. 203). Measurement of fidelity can aid all three of these recommendations. A better understanding of fidelity, methods for assessing it, and its relationship to student outcomes can help, (a) researchers to refine interventions, (b) create targeted professional development opportunities, and (c) schools understand how to document the fidelity of their interventions.

Refining Interventions

Though there are studies that relate fidelity to outcomes, it is unclear which components of the intervention lead to improved student outcomes. By relating fidelity of each intervention component and overall fidelity to outcomes, the field can get a better idea of the pieces of the intervention that have the greatest impact on student learning. It is possible that there are components of interventions that are more effective and should be emphasized while others are less important and can be eliminated. In addition, by expanding the definition of fidelity to include quality dimensions and improving the methodology with which the construct is measured, school-based practitioners can gain more insight into the appropriateness of the interventions being provided for each individual student. By accounting for quality of delivery and student engagement, the best possible interventions can be provided to each student because not only can teacher implementation be documented and improved, but educators can also respond to student engagement to ensure interventions are matched to their needs. An expanded definition of fidelity would have a similar effect in the research setting, calling attention to implementation variables beyond the surface leading to development of the most effective interventions possible.

In addition, variations in implementation often occur. Adaptations to implementation can be explicitly studied to gain a better understanding of the parameters of interventions being implemented and to potentially discover changes that may improve interventions by making them more effective, efficient, or economical (LeLaurin & Wolery, 1992; Orwin, 2000). Understanding these deviations from prescribed implementation can also help researchers to problem solve and create mechanisms for correcting problems and overcoming barriers (Yeaton & Sechrest, 1981). It is possible that deviations in implementation may be more effective for all or some students, and it is important to understand the effects of these deviations to help future students. Based on consistent study of implementation of an intervention, researchers and practitioners can create guides for the most effective implementation as delivery and changes to delivery have been documented (Mowbray et al., 2003). Understanding the ways in which interventions are implemented by documenting fidelity contributes to the understanding of different interventions and what is necessary for successful implementation (Gresham et al., 1993). Denton et al. (2003) explain that one of the reasons that research-based interventions are not implemented is because teachers do not have information about how to implement effective instructional practices, and teachers want proof that evidence-based practices benefit their students more than their current practices (Denton et al., 2003). Understanding the parameters of interventions by collecting fidelity data and relating it to outcomes on a consistent basis can help to remedy these issues.

Creating Targeted Professional Development Opportunities

Professional development opportunities are essential for schools to ensure that all of their students are successful (Denton et al., 2003). Collecting fidelity data aids in the process of determining necessary professional development. Fidelity data can highlight components of interventions for which teachers may need additional training (Gersten et al., 1982). This information can help schools to provide individual support in a "coaching" manner to teachers having difficulty implementing interventions. Contextualized feedback as provided within a "coaching" approach to collecting fidelity data can provide timely and individualized support to teachers in a less intimidating manner than other procedures (Chard & Harn, in press). In addition, when we better understand the parameters of interventions and the role of variability in implementation, these changes can either be detected and corrected early or shared with others as more effective and efficient options for improving student outcomes. Providing on-going support will improve the fidelity of interventions delivered in schools and can help schools to sustain their use of evidence-based practices. "Teachers who have the time, resources, and technical support needed to develop competence in the implementation of a program or practice are more likely to continue to use it despite obstacles such as demands on their time or changes in administration" (Denton et al., 2003, p. 207).

Helping Schools Document Fidelity

Fidelity is an important concept to the field of education. However, the definition of fidelity is unclear and methodology for assessing fidelity varies across studies. Further, because of these factors, the relationship between fidelity and student outcomes is unclear. Because of these gaps in the literature, it is uncertain how schools should proceed in documenting the fidelity of their interventions. The current study explored the relationship between fidelity, measured using direct observations, and student outcomes using an expanded definition of fidelity that focuses on surface/content dimensions of fidelity, quality/process dimensions of fidelity, and student engagement.

CHAPTER III

METHODS

The purpose of this study was to examine the effects of fidelity on kindergarten early reading outcomes using an expanded definition of fidelity that includes a focus on surface/content dimensions of fidelity, quality/process dimensions of fidelity, and student engagement. It explored the relation between the fidelity of three research-based interventions and kindergarten outcome measures of phonological awareness, alphabetic principle, word reading, and reading fluency. An analysis of existing data from the Project Optimize study (Simmons et al., 2007) was conducted. This chapter discusses the participants, setting, and interventions and provides a description of the fidelity and student outcome measures and data collection methods. The data analysis procedures are also discussed.

Participants

Students

In September of their kindergarten year, 116 students from seven elementary schools in the Pacific Northwest were screened on the Letter Naming Fluency (LNF) and Onset Recognition Fluency (OnRF) DIBELS measures (Good, Gruba & Kaminski, 2002; see description of measures to follow) and selected to participate in the study based on the following criteria: (a) they scored at or below the 25th percentile in the district on both measures (i.e., less than 11 on OnRF and less than 6 on LNF); and (b) their performance was confirmed by kindergarten teachers as being at risk for reading difficulty. Children were excluded who had (a) severe hearing or visual acuity problems or (b) were determined by school personnel to have significantly limited English proficiency. All participating kindergartners were then administered the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981) to determine their baseline level of receptive vocabulary knowledge.

Socioeconomic status, race, and gender were allowed to vary consistent with the district population from which the sample was selected. Participating children were primarily White (n = 94; 83.93%) and Latino/Hispanic (n =15; 13.39%). Two of the children were Black/African-American, and one did not specify race or ethnicity. Fifty-eight percent of the sample was male (n = 65); the mean age for students in the fall was 5 years 7 months, with a range from 5 years 0 months to 6 years 9 months.

Interventionists

Interventionists included 4 certified teachers and 24 educational assistants between 35 and 44 years of age. The typical interventionist had a high school education with some college coursework and an average of 5.7 years instructional experience in schools.

Setting

The study took place in seven schools across two districts in the Pacific Northwest. All seven participating schools received Title I funding, and the percentage of students qualifying for free- and reduced-cost lunch services ranged from 32% to 63%. In terms of overall enrollment, schools ranged from 319 to 683; time allocated for kindergarten in all schools was 2.5 hours per day.

Due to the young age of the children and the intensity of the interventions, group size was limited to five or fewer children. Each school implemented three instructional treatments. Two were researcher-developed as part of the Project Optimize field-initiated research grant while one was a commercially available, research-based intervention program. The Project Optimize developed instructional treatments were called phonological awareness with spelling instruction (PAS) and phonological awareness with storybook instruction (PASB). The research-based intervention program was based on the Sounds and Letters component of *Open Court Reading 2000* (OC; Adams et al., 2000). All of the interventions included phonologic, alphabetic, and orthographic activities so for the purposes of the current study, all three instructional treatments were treated as one intervention to examine the relation of fidelity to outcomes. The number of intervention groups per school varied depending on the number of students identified as at risk and the size of the school, with a maximum of six and a minimum of three intervention groups per school.

Independent Variable

Fidelity was measured using direct assessment methods (Gresham, 1989). The direct observations of fidelity focused on (a) implementation accuracy of components of the lesson, (b) level of delivery of the components, (c) overall quality of delivery, and (d) overall student engagement. Prior to conducting fidelity observations, training was completed and observers established a between-observer reliability of .85 or higher and interobserver agreement was collected on 20% of observations; details are discussed in the procedures section. The fidelity forms used in the study are included in Appendix B.

Total Fidelity

To assess surface/content dimensions of fidelity, critical components of each intervention were identified and operationalized, and each component was assessed using a 3-point scale. Fidelity was evaluated by observing complete instructional sessions and documenting the presence or absence of each critical component in real time. If the critical component was always demonstrated (>80% of the time) during the observation, 2 points were assigned, 1 point was given for a component that was observed most of the time (20-80% of the time), and no points were assigned if a component was not observed (<20% of the time). For the PAS intervention, daily lessons were composed of the same basic activities. The form contained the sequence of activities and each activity was broken down into the same key components. The components in the activities included: (a) used wording from script, (b) teacher corrected student mistakes, and (c) teacher leads/tests students on examples. For the PASB intervention, lessons were composed of 15 minutes of phonological awareness and alphabetic principle instruction that was the same as the PAS intervention and 15 minutes of comprehension instruction. The first part of the fidelity form was the same as that of the PAS intervention. The second part of the fidelity form contained operationalized components for each of the vocabulary and retell activities and included: (a) used wording similar to script, (b) pointed effortlessly to correct illustration while reviewing vocabulary, (c) used designated prompts for retell, and (d) gave each child similar opportunity to talk. For the OC intervention, lessons and activities were not as similar from day to day so observers wrote in the steps of the activities each day for the observed lesson and rated each step on the scale described above. The total fidelity score for all three interventions was calculated by tallying the observed components and dividing by the total possible components score. This score was documented as a percent of implementation.

Quality of Delivery

Overall quality of delivery was assessed in the same manner for each intervention. One question addressed quality of implementation. Similar to Al Otaiba and Fuchs (2006) who assessed weekly overall quality of lesson delivery, at the end of each fidelity observation, observers were to rate overall "Quality of Lesson Delivery" as high, medium, or low.

Student Engagement

Student engagement was also assessed in the same manner for each intervention. Student engagement was measured in one question. Again similar to Al Otaiba and Fuchs (2006) who addressed weekly overall student engagement, at the end of each fidelity observation, observers were to rate "Student Engagement" throughout the lesson as high, medium, or low.

Student Measures

Onset Recognition Fluency

Onset Recognition Fluency, OnRF, task is a standardized, individually administered, beginning measure of phonological awareness that assesses a child's ability to recognize and produce the initial sound in an orally presented word. The examiner presents four pictures to the child, names each picture, and then asks the child to identify (i.e., point to or say) the picture that begins with the same sound the examiner produces. The child is also asked to produce orally the onset for an orally presented word that matches one of the given pictures. The examiner calculates the amount of time taken to identify/produce the correct sound and converts the score into the number of onsets correct in a minute. Alternate form reliability of the OnRF measure is .65 and test-retest reliability ranges from .65-.90 (Good et al., 2002).

Phoneme Segmentation Fluency

The Phoneme Segmentation Fluency, PSF, task is a standardized, 1-minute, individually administered measure that assesses phonological awareness. The purpose of the PSF measure is to assess a student's ability to segment words into their individual sounds. This measure is comprised primarily of three and four phoneme words (e.g., *fish, sun*). The examiner orally presents one word at a time and the student segments the word into its individual sounds. For example, *fish* may be correctly segmented into its three sounds /f/ /i/ /sh/ to receive three points, the total possible points. Partial credit is given for the portions of the word correctly segmented. *Fish* could be segmented into two portions, /f/ /ish/, and two of the three points would have been earned. The total score is the number of correct segments produced in 1 minute. PSF has alternate-form reliability of .88 and predictive validity coefficients with other reading measures ranging from .73-.91 (Good et al., 2002).

Nonsense Word Fluency

The Nonsense Word Fluency, NWF, task is a standardized, 1-minute, individually administered measure that assesses a student's knowledge of the alphabetic principle. The purpose of the NWF measure is to assess a student's ability to produce letter-sound correspondences as quickly as possible. The measure is comprised of CVC and VC nonsense words (e.g., *rav*, *ep*). The examiner presents the student with an 8.5" x 11" sheet of paper with five different CVC/VC words per line. He/she is asked to provide the sound of each letter or read the whole word. The student is timed for 1 minute and credit is given for each letter-sound correspondence produced correctly. The student receives credit if he or she produces each individual sound or if he or she produces the entire nonsense word. For example, the nonsense word *rav* can be produced as /r/ /a/ /v/ or *rav* to receive all three possible points for that word. The total score is the number of correct letter-sound correspondences produced in 1 minute. Alternate-form reliability for NWF ranges from .67 to .87 and concurrent validity with the readiness subtests of the Woodcock-Johnson Psychoeducational Test ranges from .35 to .66 (Good et al., 2002).

Woodcock Reading Mastery Test-Revised: Word Attack Subtest (Word AT)

Word AT (Woodcock, 1987) is a standardized, individually administered test that measures a student's ability to decode a list of nonwords out of context. He or she is presented with two to six words on a page. Acceptable pronunciations are provided on the testing protocol and on the examiner's side of the display stimulus book. Administration is discontinued if the child produces six consecutive incorrect responses that end with the last item on an administered page. The test developers did not report test-retest reliability of the Word AT subtest; however, split-half reliability ranges from .91-.97. Criterion related validity of the Word AT subtest with the Woodcock-Johnson Psychoeducational Total Reading Battery for first and third grades is .69 and .68, respectively. A correlation for kindergarten was not provided.

Woodcock Reading Mastery Test-Revised: Word Identification Subtest (Word ID)

Word ID (Woodcock, 1987) is a standardized, individually administered test that measures a student's ability to read a list of real words out of context. He or she is presented with one to nine words on a page and must orally produce the correct word. Administration is discontinued if the child produces six consecutive incorrect responses that end with the last item on an administered page. Test-retest reliability of the Word ID subtest was not reported by the test developer; however, split-half reliabilities for grades 1 and 3 ranged between .97 and .99. Criterion related validity of the Word ID subtest with the Woodcock-Johnson Psychoeducational Total Reading Battery for first and third grades is .82 and .86, respectively. A correlation for kindergarten was not provided.

Oral Reading Fluency

The Oral Reading Fluency, ORF, task is a standardized, individually administered test of accuracy and fluency with connected text. The ORF passages and procedures are based on the program of research and development of Curriculum-Based Measurement of Reading by Stan Deno and colleagues at the University of Minnesota and use the procedures described in Shinn (1989). The student is presented with a grade-level passage and asked to read the passage aloud. The final score is the number of correct words read in 1 minute. Test-retest reliabilities for elementary students ranged from .92 to .97 while alternate form reliability of different reading passages drawn from the same level ranged from .89 to .94 (Tindal, Marston & Deno, 1983). Criterion-related validity studied in eight separate studies in the 1980's reported coefficients ranging from .52 to .91 (Good et al., 2002; Good & Jefferson, 1998).

Procedures

Interventionist Training

All interventionists received a two-day training prior to intervention. The researcher authors from the Project Optimize study provided training on the two researcher-developed interventions while one of the co-authors of the *Open Court Reading 2000* curriculum provided training on the commercially available program. Training focused on modeling and familiarizing the interventionists with the materials, lesson formats, and teacher wording. A large portion of time was spent practicing lessons in order for the trainers to provide feedback. Throughout the training all interventionists were observed to assess fidelity and provide feedback and support. In addition, two follow-up training sessions were held during the intervention in January and March for all interventionists to provide more training on new instructional methods, activities, and materials.

Data Collector Training

Fidelity Observations

Graduate students from the University of Oregon College of Education were recruited and trained to observe fidelity prior to the start of the study. In addition to receiving training prior to the start of the study, graduate students also met weekly with the principle investigators to discuss things that were seen in schools while conducting the fidelity observations as well as what was meant surface/content or total fidelity, quality, and engagement as the intervention evolved. Because the interventions were part of a research study, the nature of the activities within each lesson changed. Meeting on a regular basis ensured consistency in observers both within and across school sites (Hayes, Nelson, & Jarret, 1986).

Student Assessments

Graduate students from the University of Oregon College of Education were recruited and trained to administer and score all dependent measures prior to the start of the study. A member of the Project Optimize study delivered trainings. The focus of the trainings was on the administration and scoring of each measure while special attention was paid to the importance of following standardized procedures. The data collectors practiced administering the measures, asked questions, and received feedback on administration of each measure. Prior to collecting data, each data collector demonstrated at least 90% reliability for both administering and scoring.

Intervention Implementation

At-risk students were randomly assigned to one of the three interventions using a stratified random sampling procedure. Teachers and teaching assistants were randomly assigned to one of the three interventions. From November through mid-May, students received one of the three, 30-minute early reading interventions supplemental to their typical 2.5-hour kindergarten day. The small-group interventions, which occurred during extended kindergarten hours (i.e., either before or after the regular kindergarten instructional day), were conducted by either certified teachers or teaching assistants at the child's school. One interventionist led each group for the course of the study; however, two of the small groups had turnover in interventionist during the intervention. On average, children received 108 days of supplemental, small-group intervention for a total of 54 hours over the year.

Measuring Fidelity

Fidelity observations were conducted by research team members at seven points during the year: twice per month during the first two months of intervention and once every three weeks for the remaining weeks of intervention. One graduate student was assigned to each school and worked with all interventionists in the school across the entire intervention. This allowed for consistency in the observations at each school as well as a more collaborative, coaching approach. Observations were scheduled with interventionists in advance and immediately following each intervention, the observer and interventionist discussed methods for improving intervention delivery.

Measuring Dependent Variables

Dependent variables were assessed post-intervention. PSF, NWF, and the Word AT and Word ID subtests of the *Woodcock Reading Mastery Test-Revised* were administered in May at posttest only. Two ORF passages were also administered at posttest only.

Data Analysis

The design was a nested, hierarchical design. At level 1, student outcome measures (e.g., NWF) were nested within level 2 small intervention delivery groups, hereinafter referred to as simply groups. Groups were also nested within schools but because (a) the number of schools is limited and (b) no hypotheses about school effects over and above group effects were developed, school effects were not included. The fidelity measures were also repeated over time but were at the group level and in general were not timed to coincide with student level outcome assessments.

Given this, multi-process multi-level (MPML) models were used to study the interrelations among the fidelity measures and interrelations among group levels of fidelity and group levels of student outcome. The MPML model for the fidelity measures involved average levels and slopes for the 3 fidelity constructs simultaneously. The MPML model estimated the correlations among intercepts and slopes and indicated how the fidelity measures were related. These analyses were just standard 2 level growth models, repeated measures within groups and groups. Because the sample size at the group level was quite modest (n = 27) and the problems with hypothesis testing of random effects in modest samples are well known, confidence intervals (.999, .99 and .95) on the estimated parameters were examined in addition to the standard results like point estimates, critical ratios and likelihood ratio tests (Pinheiro & Bates, 2000).

The MPML model examining both fidelity and student outcome was more complicated. The primary interest was in how fidelity, either average level or slope, was related to the group level student outcome final status. Although student outcomes varied within groups as well as between groups, because the fidelity measures are only group level measures, the within group student outcome variation was not of central interest.

Estimation for relating the dimensions of fidelity to each other was carried out using LME (Pinheiro & Bates, 2000), which is implemented in both commercial S-Plus (S-Plus, 2006) and freeware R (R Development Team, 2007). Estimation for relating the dimensions of fidelity to student outcomes was carried out using Mplus (Muthén & Muthén, 2007) rather than LME because MPlus allows for structural relations among the latent variables, whereas LME does not. Although some multi-level modeling packages have explicit routines for multiple dependent variables (multi-process models), standard packages without such extensions, such as LME, can still be used by setting up a multiple strata model. In this case, however, one of the variables, fidelity, varied only at the group level and not at the student or within student level. Therefore, the standard procedure for a multiple strata (multi-process) model was altered slightly. The fidelity measures for a given group were entered for the first student in each group and set to missing for the remaining students within the group. Growth models were fit for each fidelity and outcome measure separately as a prelude to fitting multi-process models. Once the separate models were adequate, the models were combined into a multi-process model.

CHAPTER IV

RESULTS

This study explored the concept of fidelity to aid both researchers and practitioners in their measurement of the construct and use of the data. This study examined how (a) the dimensions of fidelity relate to each other and (b) the dimensions of fidelity relate to student early literacy outcomes. Multi-process multi-level (MPML) models were used to study the interrelations among the dimensions of fidelity and the interrelations among the group level fidelity measures and group level measures of student outcome. The results of these analyses are presented here, beginning with general preliminary analyses and descriptive statistics and then results related to the primary research questions for the study are presented. Tables are presented in Appendix C and Figures in Appendix D; each are presented sequentially.

Preliminary Analyses and Descriptive Statistics

Assignment of observers to groups and observations was first explored. Each observer observed 3 or 4 groups consistently throughout the year. With two exceptions, observers and schools were confounded (i.e., one observer was assigned to each school and observed all interventionists at that school all year). The number of fidelity assessments by group and time period was examined. Ratings of quality and engagement were more often missing then the total fidelity measure. Group 6, for example, was missing all quality and engagement ratings. Also worth noting is that fewer observations were done in the early period (October-December; average of 2.9) than in the late period (January-May; average of 4.0). The number of observations has a direct bearing on how reliable any aggregate index of fidelity will be for a given time period, with more assessments leading to higher reliability (Gersten et al., 1982; Stoolmiller et al., 2000).

Descriptive statistics for the measures are shown in Table C1. At the observation level the total fidelity measure had a strong ceiling effect, many scores at the maximum value, and a strong negative skew. To make the total fidelity measure more amenable to standard methods, the scale was reversed so that it was total infidelity, and it was square root transformed to reduce skewness. For the quality ratings, 2 scores of 1 were trimmed to 2 to reduce the potential impact of these outliers. The quality and engagement ratings were very coarse (few distinct values) and no transformation will render such distributions normal. The trimmed and transformed versions of the fidelity constructs are shown in a scatter plot matrix in Figure D1. Each plot has a fitted linear regression (dashed line) and a non-parametric smooth regression (solid line) to check for nonlinear trends. Regression statistics are shown in the top margins of the scatter plots (r = correlation, b = regression weight, t = t statistic for regression weight, p = p level for t statistic, N = sample size) and descriptive statistics are shown in the top margins of the normal quantile plots along the main diagonal. Regression statistics are suggestive but should not be taken too definitively because they are based on the unlikely assumption that the repeated observations on a group are independent. Quality is more strongly related to total infidelity ($r = -.43$) than engagement ($r = -.21$). Quality and engagement are moderately correlated ($r = .45$). When total infidelity was plotted as the dependent variable, relations with quality and engagement appeared fairly linear. The coarseness of the quality and engagement distributions made it difficult to interpret the plots when they were plotted as the dependent variables against total infidelity.

Growth curves for the fidelity measures for all 27 groups were explored and are included in Figures D2-D4. Each plot has a fitted linear growth curve (dashed line) based on the ordinary least squares (OLS) regress of fidelity score on time (in days). Most of the plots show a large amount of occasion-to-occasion variability, which suggests that a single observation is not likely to be very reliable for any of the fidelity constructs. This finding is similar to those of other behavioral observation studies (Stoolmiller et al., 2000).

Research Question 1: Relating Dimensions of Fidelity

Individual Fidelity Growth Models

Growth models were developed for each of the fidelity dimensions: engagement, quality, and total infidelity. For each dimension of fidelity a consistent approach was applied to obtain the most parsimonious model of the data. First, a random slope, a random intercept, and the correlation between the two were included in the model. In the next model, the correlation between the random intercept and slope was forced to zero. The final model contains only a random intercept.

Engagement

Results for the 3 growth models for the Engagement ratings are shown in Table C2. The first model had a random slope, a random intercept (fall status) and the correlation between the two. The second model forced the correlation between the random intercept and slope to zero, and the final model had just a random intercept. The 3 models were nested and likelihood ratio tests are shown at the bottom of the table. As is apparent, removing the random slope from the model did not significantly degrade the fit of the model to the data (model 1 vs. 3, $\chi^2 = 1.50$, $df = 2$, $p = .47$). The model estimates for the intercept only model (model 3) indicated that the fixed effect for the linear trend was also not significant. This model suggests that the groups did differ in terms of the average level of student engagement but these differences were stable over time. Consistent with the growth curve plots that showed a lot of time-to-time variability, the random intercept accounted for only 25% of the total variance. A single observation of student engagement therefore would have a very low reliability of .25 and an aggregate score based on 7 observations (the average across all groups) would have a reliability of about .70, which is still below the commonly recommended standard of .80 for regression analysis (Cohen & Cohen, 1983).

Quality

Results for the 3 growth models for the Quality ratings are shown in Table C3. The first model had a random slope, a random intercept (fall status) and the correlation between the two. The second model forced the correlation between the random intercept and slope to zero, and the final model had just a random intercept. The 3 models were nested and likelihood ratio tests are shown at the bottom of the table. Removing the random slope from the model did not significantly degrade the fit of the model to the data ($\chi^2 = 5.45$, $df = 2$, $p = .07$). However, removing the correlation between the slope and intercept did significantly degrade the fit of the model ($\chi^2 = 5.45$, $df = 1$, $p = .02$). Once the correlation was out of the model, however, removing the random slope completely (i.e., removing the standard deviation) did not further significantly degrade the model ($\chi^2 = .00$, $df = 1$, $p = 1.0$). To explore this issue a bit more, the quality ratings were reversed and recoded so that the original values of 3, 2.5 and 2 were set equal to 0, 1 and 2 respectively. With this set of values, the quality ratings now resembled a count variable that takes on integer values, an outcome which is more typically modeled with a poisson regression. When the same set of 3 models described above were re-specified as multi-level models for count (poisson) data, there was no support for a random slope ($\chi^2 = 0.62$, $df = 2$, $p = 0.73$), which along with the principle of parsimony tends to support model 3 with no random slope.

In model 3, the fixed effect for the slope was positive and significant indicating an upward trend in quality ratings over time. Consistent with the growth curve plots that showed a lot of time-to-time variability, the random intercept accounted for only 45% of the total variance. A single observation of quality therefore would have a very low reliability of .45 but an aggregate score based on 7 observations (the average across all groups) would have a reliability of about .85, which is above the commonly recommended standard of .80 for regression analysis (Cohen & Cohen, 1983).

Total Infidelity

Results for the 3 growth models for the Total Infidelity ratings are shown in Table C4. The first model had a random slope, a random intercept (fall status) and the correlation between the two. The second model forced the correlation between the random intercept and slope to zero, and the final model had just a random intercept. The 3 models were nested and likelihood ratio tests are shown at the bottom of the table. As is apparent, removing the random slope from the model did not significantly degrade the fit of the model to the data ($\chi^2 = .85$, $df = 2$, $p = .66$). The model estimates for the intercept only model (model 3) indicated that the fixed effect for the linear trend was negative and significant indicating a drop in infidelity or equivalently an increase in fidelity over time. The time trend for fidelity (increasing) tended to mirror the time trend for quality. Consistent with the growth curve plots that showed a lot of time-to-time variability, the random intercept accounted for only 27% of the total variance. A single observation of infidelity therefore would have a very low reliability of .27 and an aggregate score based on 7 observations (the average across all groups) would have a reliability of about .72, which is still below the commonly recommended standard for regression analysis (Cohen & Cohen, 1983).

Multi-Process Fidelity Model

Because the sample size at the group level was quite modest (n = 27) and there are problems with hypothesis testing random effects in modest samples, confidence intervals (.999, .99 and .95) on the estimated parameters were examined in addition to the standard results like point estimates, critical ratios, and likelihood ratio tests. Table C5 shows the estimates and their associated significance level and likelihood ratio tests for 4 models, labeled model 1 to model 4. Model 1 was the most general and includes all 3 possible correlations among the random intercepts for Quality, Total Infidelity, and Engagement. Model 4 was the most restricted and did not include any correlations. As is apparent, the likelihood ratio test for model 1 vs. 4 indicated that eliminating all possible correlations significantly degraded the fit of the model ($\chi^2 = 17.91$, $df = 3$, $p < .001$) indicating that at least 1 correlation was significant. Model 2 eliminated the Infidelity-Engagement correlation, which did not appear to be significant in model 1 based on the point estimate and critical ratio (and also based on the confidence interval approach), and this did not significantly degrade the fit of the model ($\chi^2 = 2.81$, $df = 1$, $p = .094$). Once the Infidelity-Engagement correlation was removed, the Quality-Engagement correlation in model 2 was no longer significant (also verified by the confidence interval approach) and was eliminated in model 3. The comparison of model 2 to 3 indicated that this did not significantly degrade the fit of the model ($\chi^2 = 3.18$, $df = 1$, $p = .074$). Finally, eliminating the Infidelity-Quality correlation did significantly degrade the fit of the model as evidenced by the comparison of model 3 vs. 4 ($\chi^2 = 11.916$, $df = 1$, $p = .001$). Thus, it would appear that the only strongly significant correlation was the Infidelity-Quality correlation of about -.80. Model 1, however, did suggest that the Quality-Engagement correlation was also significant at about .69. The results are not entirely clear or consistent with respect to the Quality-Engagement correlation.

Research Question 2: Relating Dimensions of Fidelity to Outcomes

Table C6 shows estimated means and variances at the within and between group level for each of the May outcomes (Word ID, Word Attack, NWF, PSF and the two different ORF passages, ORF1 and ORF2). Column 5 also shows within and between variance proportions (ratio of either within or between variance to total variance, the sum of the within and between variances). All of the within group variances were strongly significant and constitute from 68 to 91 percent of the total variance. Results for between group variance were not as clear. Only ORF1 had a significant critical ratio but both ORF1 and PSF had significant between group variance using a nested chi-square test (comparing the model with no between group variance to one with freely estimated between group variance). The difficulty of significance testing of variance components is well known and was discussed previously. The nested chi-square test is known to be conservative in the sense that one is likely to conclude that there is no variance based on the *p* value of the test when in fact there is. In other words, the true *p* value is actually smaller than the computed *p* value based on the chi-square distribution. The known bias suggested that ORF1 and PSF probably both show significant group level variance. Results for the other outcomes were less clear although the between variance proportions tended to be smaller, .14 or less, compared to ORF1 and PSF.

Fidelity and May Outcome Models

Tables C7-C9 show results for the fidelity constructs predicting May outcomes. Because the outcome in each model was a single measure in May, there was no outcome slope to predict and the only student level effect was the outcome within group variance, which is in the first line of each table. The rest of the effects were group level effects. In using the observation data, the repeated measures for the fidelity constructs were aggregated across time to produce 4 indicators, 2 in the fall and 2 in the spring and then these individual indicators were grand mean centered about zero to reduce the possibility of convergence problems. Thus, the mean level of each fidelity construct was about zero in Tables C7-C9. All estimation was carried out using Mplus because it allows for structural relations among the latent variables, whereas LME does not.

Infidelity Predicting May Outcome

The Infidelity effect on outcome line of Table C7 shows the effects of infidelity on the May outcomes. As can be seen, the effects were all positive, opposite of what was hypothesized, although only in the case of Word Attack was the effect significant ($.01 < p < .05$). The residual between group outcome variances were all non-significant. This could be because all of the between group variance is accounted for by infidelity. However, the lack of significant effects in Table C7 and the fact that in Table C6, most of the between group variance estimates were non-significant even without infidelity in the model suggest that there was not much between group variance to predict. The last 2 rows of Table C7 show the proportion of outcome variance at the group level and the group level *R* squared, that is, the proportion of group level variance in the outcome accounted for by infidelity. The *R* squared statistics were substantial but mostly for those outcomes where it is questionable whether there was any group level variance. For Word Attack, the only outcome with a significant effect, the *R* squared was .81.

The second and third lines of Table C7 give the infidelity residual and construct variances, 1.43 and .64 respectively, which imply a reliability of a single aggregate indicator of infidelity of about .31. These values were similar to the corresponding values in Table C4 after transforming standard deviations to variances (1.81 and .66) and as noted previously a single observation had a reliability of .27. The spring indicators of infidelity were allowed to have a common intercept, -.66, which was significant at $.001 < p < .01$, to model the decrease in infidelity between fall and spring. The infidelity mean was close to zero because the indicators were all grand mean centered to prevent convergence problems. The outcome intercept was the mean level of the outcome when infidelity is zero, which because of the centering of the infidelity indicators was about the average level of infidelity.

Engagement Predicting May Outcome

The Engagement mean effect on outcome line of Table C8 shows the effects of engagement on the May outcomes. As can be seen, the effects were all positive as hypothesized although only in the case of ORF (1 and 2) were the effects significant. The residual between group outcome variances were all non-significant. This could be because all of the between group variance was accounted for by engagement but the lack of significant effects in Table C8 and the fact that, in Table C6, most of the between group variance estimates are non-significant even without engagement in the model suggested that there is not much between group variance to predict. The last 2 rows of Table C8 show the proportion of outcome variance at the group level and the group level *R* squared, that is, the proportion of group level variance in the outcome accounted for by engagement. The *R* squared statistics were substantial but mostly for those outcomes where it was questionable whether there was any group level variance. For ORF 1 and 2, the only outcomes with significant effects, the *R* squared statistics were .38 and .59, respectively.

The second and third lines of Table C8 give the engagement residual and construct variances, .07 and .05 respectively, which imply a reliability of a single aggregate indicator of engagement of about .42. These values were somewhat higher than the corresponding values in Table C2 after transforming standard deviations to variances (.11 and .04) and as noted previously a single observation had a reliability of .25. It would appear that aggregation enhanced the reliability of engagement more than infidelity, which was perhaps not surprising given the crudeness of the engagement ratings. The engagement mean was close to zero because the indicators were all grand mean centered to prevent convergence problems. The outcome intercept was the mean level of the outcome when engagement was zero, which because of the centering of the engagement indicators was about the average level of engagement.

Quality Predicting May Outcome

Unlike the infidelity and engagement models, the quality models in Table C9 included the quality slope. There was enough ambiguity about the importance of the quality slope (Table C3) to make it seem worth the effort to include it, at least to begin with, as a predictor of May outcome. The quality effects in Table C9 were uniformly positive as hypothesized but none were significant. The residual between group outcome variances were all fixed to zero. All of the models had serious convergence problems if the group level residual variance was estimated so it was fixed to zero. The last 2 rows of Table C9 show the proportion of outcome variance at the group level and the group level *R* squared, that is, the proportion of group level variance in the outcome accounted for by quality. Because the group level residual variance was fixed at zero, the *R* squared statistics were all 1.0 and not particularly meaningful. The proportion group variance was also not particularly meaningful given the problems with the models.

To better understand the problems with the quality models, Table C9 shows the correlation between the quality mean level and slope for each model and as is apparent, it was significantly different from zero and very close to -1 or even greater than -1 for the PSF model. As noted previously in Table C3, if the correlation was removed from the model, the fit was significantly degraded although if the slope was then removed completely in a second step, the fit was not further significantly degraded. One possibility not discussed previously is that this strong negative correlation was a consequence of the coaching of interventionists that went with the observations and the crude nature of the quality ratings. In other words, if the initial quality was observed to be poor, the observers tried to coach the interventionists to improve. If this coaching was successful, it would have the effect of creating a strong negative correlation between initial quality ratings and change over time in quality ratings because initially low scoring interventionists would improve. This would not affect initially high scoring interventionists because they would not get coaching and would have little room for improvement on the crude scale even if they did.

Given that the strong negative correlation may actually represent a substantively interesting phenomenon and not just a statistical artifact, a second version of the quality model was estimated. The model was re-parameterized to represent initial quality and slope of quality and initial quality predicted the quality slope perfectly, that is, the slope residual variance was set to zero. Because the quality slope is perfectly predicted by initial quality, only one of these two can be used to predict May outcome and the choice is arbitrary. Whatever results are obtained for one will be the same as using the other except with opposite sign because initial status is negatively related to slope.

The second version of the quality model is shown in Table C10. All of the effects of the quality slope on May outcome were positive as hypothesized but not significant and substantially smaller than the estimates in Table C9 that were inflated by the collinearity between average level and slope. The *R* squared statistics are also small compared to results for infidelity and engagement.

CHAPTER V

DISCUSSION

Fidelity has become an important topic in the field of education for both research and instructional practice, especially with its inclusion within RTI methodology in IDEA 2004. The definition of fidelity has been expanded in several fields to encompass quality of intervention delivery and student engagement in addition to the traditional focus on delivery of surface or component level features. Although leaders in the field of education have highlighted the need to consider these additional dimensions of fidelity, the issue still poses a challenge three reasons: (a) there is limited agreement on a definition of fidelity, (b) varying methods are used to measure the construct, and (c) inconsistent relations between fidelity and outcomes have been demonstrated. The purpose of the current study was to address these gaps and develop a greater understanding of fidelity by examining its multiple dimensions and their relation to student outcomes.

To do so, the following definition of fidelity was used: *the degree to which central surface level intervention components are implemented with quality such that students are engaged in the intervention.* Based on this definition the following research questions were addressed: What is the relation between dimensions of fidelity (total/surface fidelity, quality of delivery, student engagement)? What is the relation between dimensions of fidelity and student outcomes measured using multiple early literacy measures? A brief summary of the results will be provided next, followed by limitations and implications for both research and school practice and finally potential next steps to further our understanding.

Summary of Results

Change and Reliability of Dimensions Across Time

The results of the current study indicate that overall, there was a large amount of variability in all three of the fidelity measures (total fidelity, engagement, and quality) which indicates that a single observation is not highly reliable or predictive of future measurement. Reliability coefficients for single observations for all three dimensions of fidelity ranged from .25 to .45 and from .70 to .85 when all 7 observations for each fidelity measure were aggregated across time. When examining each dimension of fidelity across the duration of the intervention, student engagement varied across intervention groups, but each group demonstrated similar levels of engagement across time. The dimensions of quality of delivery and total fidelity (surface level) each increased across the duration of the intervention. The next section will discuss the complex relations of these fidelity dimensions to each other.

Research Question 1: Relating Dimensions of Fidelity

The results demonstrated that the construct of fidelity is indeed multidimensional and potentially more complicated than researchers had considered (Gersten et al., 2005; Moncher & Prinz, 1991; Power et al., 2005). In terms of relating the dimensions of fidelity to each other, total fidelity and quality were significantly related (r =.80), quality and engagement may be significantly related (r = .69) as results of the multi-process fidelity models were not consistent, and total fidelity and engagement were not significantly related (r = .49).

Research Question 2: Relating Dimensions of Fidelity to Outcomes

It was hypothesized that average level or slope or both for group level fidelity would be associated with end-of-year student performance. In other words, it was expected that those groups that have the highest fidelity scores averaged over time or the highest improvement in fidelity (slopes) over time would also have students that performed the highest at the end of the intervention. The amount of variability within each group on the outcome measures was greater than the amount of variability between groups. Therefore, there was not a lot of between group variance to account for which, coupled with the small sample size, impacts the conclusiveness of these results.

The relation between average total fidelity and student outcomes was in the *opposite* direction of what was hypothesized—lower total fidelity was related to higher student outcomes. This relation was significant for only one of the student outcome measures (Word Attack). The relation between average student engagement and student outcomes was in the hypothesized direction—higher engagement was related to higher student outcomes. This relation was significant only for oral reading fluency. Finally, the relation between quality of delivery and student outcomes was more complex. Change in quality over time (slope) was included as a predictor in the model in addition to average quality because the individual quality growth model indicated that the change in quality needed to be explored. The relation between average quality and/or change in quality over time with student outcomes was also in the hypothesized direction with higher quality related to higher student outcoimes. However, quality was not significantly related to any of the student outcomes. When initial quality was set to predict change in quality over time to model the possible effect of coaching (consultation with interventionists to improve implementation) on implementation for interventionists who started with lower fidelity, the effects were also positive but not significant.

Although the results from this study did not always align with the hypotheses and were not conclusive, they highlight several issues related to fidelity that need to be considered by both researchers and practitioners in the field of education. The remainder of this chapter is divided into three sections. First, limitations of the current study are discussed followed by implications for researchers and school practitioners related to defining and measuring fidelity and its relation to student outcomes. Finally, future directions for research and school practice are highlighted.

Limitations

Though the findings provide insight into the concept of fidelity, they must be considered in light of several limitations, some of which have been mentioned. To begin, this analysis was conducted retrospectively. The initial study, described in Simmons et al. (2007), was an intervention study focusing on variables to improve student outcomes and not a study designed to understand the facets of fidelity related to such outcomes. Consequently, there was limited variability in student outcome performance. This is not surprising considering students in the study were chosen to participate based on their significantly low early reading skills. Results may have looked different if there had been more variability in student outcomes to predict. The methods for conducting the fidelity observations, while more robust than most studies capture, had significant limitations. Specifically, the student engagement and quality of implementation scores were a simple 1-3 rating, which caused significant challenges during analysis due to the restricted range and the nonnormal distribution of the data. Additionally, there was a fair amount of missing engagement and quality of delivery data which must be considered in interpreting the results. The small number of groups and observations (which impacted the reliability of the fidelity measures) may have also restricted our ability to unpack the dimensions of fidelity more precisely.

In addition, fidelity observations were set up to be conducted as part of a coaching model. One coach was assigned to a school and was responsible for collecting all observations for each interventionist across the duration of the study as well as to coach/consult with them to improve the quality of implementation and fidelity. This, of course, impacted the level of fidelity of the interventionists. In general, the initial observations indicated a fairly high average level of implementation, providing a ceiling effect, but fidelity only continued to improve across the duration of the intervention. As a result, fidelity was not allowed to vary as it would naturally which may have impacted these results and may be necessary for conducting research on fidelity (O'Donnell, 2008). Also, because of this model, observers and schools were totally confounded. One observer was assigned to a school making it difficult to determine whether fidelity scores were dependent on the observer or the school context.

Another important limitation that should be considered in future studies, was the process for conducting the observations. Within the same observation, and by the same observer, all three dimensions of fidelity were collected, meaning that they are not independent of one another. The observer first watched the entire lesson checking off the components related to the total fidelity score and then rated the session on quality of delivery and student engagement after having observed the entire lesson. Therefore, the ratings of quality and engagement may be totally confounded with the total fidelity score. Findings may have looked different if quality and engagement had been rated independently by a separate observer who had not processed the specific steps of each lesson.

Implications for Research

By measuring fidelity and systematically studying how it relates to outcomes, researchers may be able to identify the most essential facets of interventions to make interventions more efficient and effective. The typical approach of assuming that if an intervention is implemented with fidelity, it is the cause of improved outcomes is debatable given these findings as well as others (Al Otaiba & Fuchs, 2006; Dane & Schneider, 1998; van otterloo et al., 2006; Zvoch et al., 2007). Within research studies, there are many threats to internal validity, so documenting and empirically demonstrating that measured fidelity is related to outcomes helps to demonstrate the causal relation between the intervention (independent variable) and improved outcomes (Peterson et al, 1982). The remainder of this section will discuss implications of this study to research by focusing on issues related to fidelity measurement, how the data will be used in research, and the relation of fidelity to outcomes.

Measurement Issues

These results point to the need for researchers to continue to explore and analyze the construct of fidelity as well as methods for measuring fidelity and collecting the data. We found that these measures of fidelity were not highly reliable, not at the inter-observer level (the typical reliability facet examined) but for a given interventionist across time. This interventionist variability should be considered during intervention studies to ensure that enough observations are conducted to get an accurate picture of day-to-day implementation. Similar reliability issues have been identified in studies preventive interventions for conduct disorder and child aggression (Stoolmiller et al., 2000).

An additional issue that may impact how fidelity is measured in research is how the differing dimensions were related to each other. In the current study, quality and fidelity were significantly correlated while engagement and fidelity were not. Both quality and engagement were measured using one question at the end of an observation. Researchers need to continue to explore ways of measuring fidelity so that we are getting the most useful and meaningful information possible. Though these results are inconclusive, it appears that including overt measures of quality of delivery and student engagement may provide researchers with additional information and supports an expanded definition of fidelity. Though the relationships were not significant, quality of delivery and engagement were positively associated with student outcomes. The additional focus on quality and/or engagement helps researchers to consider student responsiveness and overall quality of delivery in intervention studies and to study how adaptations may impact a program. This approach may assist in closing the research-to-practice gap by developing a more overt reciprocal relation between schools and researchers which may aid in refining interventions in meaningful ways (Denton et al., 2004; Klingner, 2004; LeLaurin & Wolery, 1992; Orwin, 2000).

Purpose of Fidelity Collection

Researchers need to consider the purpose for measuring fidelity in their studies. The type of information that a researcher is trying to obtain has a direct impact on how fidelity is defined and measured. O'Donnell (2008) discusses this issue by delineating approaches to be used in *efficacy* or *effectiveness* studies. In an efficacy study run in highly controlled conditions, it is important to have clear control over fidelity of implementation as the purpose of such studies is to document if an intervention delivered as designed and packaged is impacting student outcomes. On the other hand, in an effectiveness trial in more naturalistic settings (i.e., during intervention development and piloting in schools), it is important to allow fidelity to vary and study those variations in implementation and their impact on outcomes to better understand interventions in the real world setting and to assist with dissemination of scientifically based practices (Smith, Daunic, & Taylor, 2007). This approach is broader than the typical approach of simply documenting whether or not the intervention was implemented (i.e., surface/component or total fidelity). However, considering that in this study total fidelity and quality were highly correlated, researchers may be able to simply measure a general indicator or overall rating of implementation to document whether or not it was implemented. Future research should address whether an observation of component fidelity and an independent rating of quality are indeed correlated. If a researcher is trying to study changes to implementation or specific components of an intervention and how they impact outcomes, it will be necessary to assess the surface/component level of fidelity and relate it to outcomes. Again, if researchers want a full picture of an intervention and want to better understand how quality of delivery and student engagement impacts student outcomes, the fidelity approach should include these dimensions.

Relating Fidelity to Outcomes

The consideration of efficacy versus effectiveness brings up the debate of having rigorous standards for fidelity versus advocating for the allowance of adaptations to evidence-based practices in applied settings. The field of education is advocating the implementation of evidence-based interventions. As previously discussed, the field knows a lot about evidence-based interventions, but it is a known fact that they aren't being implemented in schools in high numbers (Denton et al., 2003; Vaughn & Damman, 2001). Through future studies that examine this issue, we can understand what level of implementation is necessary to achieve optimal outcomes for students. For example, if an intervention can be implemented with moderate fidelity and get the same outcomes as the same intervention delivered with high fidelity, that impacts what is considered "good" or "bad" fidelity as well as professional development and coaching. We need to strike an appropriate balance between advocating rigorous standards for fidelity and considering the implementation of evidence-based interventions in practice (Dane & Schneider, 1998; Leventhal & Friedman, 2004; Power et al., 2005). Adaptations made to evidence-based practices based on the needs of the students may actually provide for a more effective intervention (Castro, Barrera, & Martinez, 2004; Perepletchikova & Kazdin, 2005; Zvoch et al., 2007).

Anecdotally, interventionists in this study were highly responsive to the needs of their students. It is possible that changes they made to implementation, while giving them lower fidelity scores, actually improved student outcomes as they were responding to the needs of individual students. For example, one interventionist was a highly trained special education teacher and after working with her students for a period of time, she independently determined which skills the group had mastered and skipped ahead in the lesson to maximize student learning. The approach, while negatively evaluated in typical fidelity approaches, misses the nuance of quality instruction and effective teaching that have been demonstrated for decades (Brophy & Good, 1986). Further study of the relationships between these three dimensions of fidelity and student outcomes in effectiveness studies will improve our understanding of these issues.

These results also point to the need for further investigation of the relation of fidelity to student outcomes using appropriate data analytic methods. Fidelity data is inherently multi-leveled in structure (Zvoch et al., 2007). Appropriate data analysis should take this into account and consider that fidelity measures are collected at the group or teacher level while outcome measures are collected at the student level. The current study used complex MPML models to address these and other issues. Again, because of the nature of the data and the limitations previously discussed, the results were inconclusive but highlight the need for researchers to continue to study these relationships.

Implications for School Practice

Findings highlight several issues for the field of education to consider when recommending the measurement of fidelity in school settings. These issues will be presented in a similar fashion as the research implications but will focus on the implications for school practice. Implications related to measurement, purpose of fidelity data collection, and using fidelity data to improve student outcomes, the main focus of schools, will be discussed.

Measurement Issues

The relations between the different dimensions of fidelity have direct implications for defining the construct. For example, fidelity and quality were significantly correlated. This may indicate that when observing component/surface level fidelity and quality of delivery we are measuring similar variables. Considering that our measure of quality of delivery was simply a rating on a 3-point scale, the ease and directness of such an approach is appealing in school settings. Additionally, we should consider how the fidelity information will be communicated to teachers and make it meaningful. Providing feedback on the quality of delivery may be more meaningful than saying that the intervention was implemented with 85% accuracy. Related to this we need to consider who will collect the data and how the fidelity data will be collected.

Collecting fidelity data is incredibly resource intensive. Furthermore, policies and procedures for how to do this have not been widely disseminated (Batsche et al., 2006; NJCLD, 2005; Zvoch et al., 2007). Because there is no agreement on the definition of the construct as well as methods for data collection, schools are left to determine how to measure fidelity on their own. Guidance on a parsimonious and useful approach is critical. For example, in this study, observers were provided hours of initial training to get reliable before collecting such data and then had to schedule times for the observations across the school year. Who in the school will collect such data, provide training/oversight, and devote time to this practice across the year? Once we have identified the person to collect, the next question is, and most relevant to this study, how do we measure fidelity?

The typical approach in research, and now in schools, is to focus on component/surface level of fidelity. When collecting surface level fidelity data, checklists need to be developed for each individual intervention, which requires additional expertise and time and makes generalizations across interventions challenging. The finding that quality of implementation was significantly related to the overall component fidelity may indicate this level of specificity is neither necessary nor helpful. By expanding the definition of fidelity to include quality of delivery and engagement, we not only collect more useful information but also highlight the need to focus on delivering interventions well and in consideration of students' needs in addition to adherence to a protocol. If these results are replicated and an overall rating of quality of implementation is highly correlated with total fidelity data, it is possible that brief observations that focus on quality of delivery can be used to document fidelity more readily and be tailored to each individual context to aid in providing coaching and professional development to educators. Gersten et al. (1982) also found that their objective DISC measure correlated at a level of .45 with an overall global rating of implementation.

Another measurement issue identified in this study was the reliability of individual and aggregated observations, which has been an issue in previous studies (e.g., Gertsten et al., 1982; van Otterloo et al., 2006; Zvoch et al., 2007). Low reliability estimates were found across all dimensions of fidelity. One observation is probably not going to provide and accurate picture of day-to-day intervention delivery, yet this is most likely the typical practice in schools due to time and personnel. In the current analysis, significant variability across each observation was found; it is unclear how much of this was related only to the interventionists or possibly the interventionist-student interaction (Perepletchikova & Kazdin, 2005; Stoolmiller et al., 2000). Considering the typical variability of kindergarteners, the target students for this study, the impact of how an interventionist responds to the mood of the group should be considered and may impact the number of observations necessary for reliably determining the level of implementation. For example, this study found that 7 observations were necessary before a reliable estimate of quality of delivery was determined. Contrastingly, 7 observations of surface fidelity and student engagement did not improve the estimate to acceptable levels.

This finding significantly impacts what schools may need to do if fidelity is going to be measured with a level of certainty or rigor.

Role of Fidelity in School

As previously discussed, conducting fidelity observations can be useful in helping to bridge the research to practice gap through providing professional development through coaching (Chard & Harn, in press; Gersten et al., 1982). This study provides potential evidence that a coaching model can help to improve fidelity. Results indicated that the average level of fidelity improved across the duration of the study. The interventionists in this study, most of whom were paraprofessionals, received extensive upfront training but still had improved fidelity scores over time in this coaching context. Even though interventionists were provided extensive training up front, we cannot assume that initial success or competence with a curriculum or program indicates long lasting success (Dobson & Singer, 2005). The results of the quality model indicate that coaching may have had a significant impact on change in quality over time. When initial level of quality was correlated with change in quality over time to model whether interventionists who had lower fidelity scores initially had more change in quality over time, the results were in the hypothesized direction but not significant. This may have been because of the low sample size and future research should address this issue.

Relating Fidelity to Improved Outcomes

In practice, we also need to begin considering the appropriateness of the interventions we are providing. The field has pointed out the need to implement evidence-based practices. However, we need to move beyond this to consider which evidence-based practice is appropriate for an individual student and to provide the coaching and support needed to implement these interventions (Chard & Harn, in press; Jones, Wickstrom, & Friman, 1997). By examining quality of delivery and student engagement in addition to surface level fidelity, educators can be more responsive to students needs leading to the most effective interventions possible.

Conclusions and Potential Next Steps

The current study has highlighted the need for further study of the construct of fidelity and its relation to student outcomes. School-based practitioners must consider their purpose for measuring fidelity which will then inform their procedures in terms of what to focus on in the fidelity observation as well as how often it must be assessed. Researchers must continue to examine multiple dimensions when assessing fidelity to better understand the nuances of fidelity and how it can inform intervention refinement and potentially be more meaningful to schools. To do so, additional replication of the current findings will need to be demonstrated and studies that specifically examine fidelity will need to be conducted. As a field, we must understand fidelity and how different dimensions of fidelity relate to each other and to outcomes before fully advocating measuring the construct in practice.

APPENDIX A

TABLE OF EMPIRICAL EXAMPLES OF RELATING FIDELITY TO STUDENT OUTCOMES

Table A1

Empirical Examples of Relating Fidelity to Student Outcomes

Article	Method Used to Measure Fidelity		Results
	Surface/Content	Quality/Process	
Witt, Noell, LaFleur, & Mortenson, 1997	Permanent products were collected and fidelity was calculated as the percentage of correct permanent products received divided by the total number of treatment steps for the day.	N/A	Student academic performance increased for 3 out of 4 students when fidelity was higher.
Persampieri, Gortmaker, Daly, Sheridan, & McCurdy, 2006	Study 1: Sessions were recorded on an audiotape. S researcher listened to 40% of the sessions to calculate the number of steps completed and divided that by the total number of steps on the intervention protocol. A sticker reward chart served as a measure of how often the intervention was implemented. Study 2: Parents were given a fifteen-step protocol and asked to record each step that was implemented. Parent lead sessions were audiotaped and reviewed by a researcher.	N/A	For 3 subjects, correct words read per minute decreased during weeks when integrity was low.

Table A1 (continued)

van Otterloo, van der Leij, & Veldkamp, 2006	Daily logs were filled out by parent implementers; parents recorded the lesson components completed and any problems encountered.	One videotaped home visit using five 5-point Likert scales focused on parent-child interactions including parents' level of support and child response was conducted. All scales loaded on one factor: quality of administration.	Regression analyses showed that quantity and quality accounted for 43% of the variance on post-test scores of pre-reading skills. Quantity predicted more of the variance than quality.
Al Otaiba & Fuchs, 2006	K-PALS and 1st grade PALS: Implementation was evaluated five times across kindergarten and first grade. Researchers observed three student pairs that were randomly chosen using a checklist that scored behavior as demonstrated, not demonstrated, or not applicable. An overall classroom score was created by combining the teacher and average student scores from the observation. Each student in the nonresponder study was observed once. Ladders to Literacy: Teacher calendars were used to determine the number of activities conducted.	Ladders to Literacy: Teachers were observed and given a weekly global 1 (poor) to 3 (excellent) rating addressing lesson clarity, how well the teacher's instruction fit the intent of the lesson, and the degree to which all students were engaged.	ANOVAS were conducted to determine the relationship between student responsiveness to intervention and fidelity of implementation. Statistically significant differences in implementation of Ladders and not PALS were found in relation to student responsiveness status (nonresponsive vs. sometimes responsive vs. always responsive). Post hoc pairwise comparisons using Tukey HSD showed that nonresponsive students were in classrooms where K Ladders activities were implemented with lower quality. Both nonresponsive and sometimes responsive students were in classrooms with lower fidelity for 1st grade PALS in the fall than always responsive students.

Table A1 (continued)

Gettinger & Stoiber, 2006	Record forms were used by implementers as a self-assessment. The same forms were coded by observers. On the record forms, each step of the FACET problem-solving program was broken down into 5 to 8 activities that were coded as 0 (not completed), 1 (completed, with minimum specificity), 2 (completed, with sufficient specificity).	N/A	Correlations between fidelity of implementation of each component of FACET and improvement in student behavior ranged from .47 to .77. The correlation between fidelity and grade level was -.46. The program was implemented with higher fidelity for younger children, and children in younger grades also made greater gains in positive behavior.
Telzrow, McNamara, & Hollinger, 2000	Fidelity was measured using two work products. The first was a problem-solving worksheet that listed all of the problem-solving components; the second was an evaluation team report form.	A Likert scale and scoring rubric were used to evaluate the work products, focusing on implementation of the problem-solving components and student outcomes.	Fidelity ratings for six of the eight problem-solving components were modestly significantly correlated with ratings of outcomes. The two components with the lowest fidelity ratings were not significantly correlated with outcomes. A stepwise multiple regression analysis showed that two problem-solving components were significant predictors and accounted for 8% of the variance.

Table A1 (continued)

Kovaleski, Gickling, Morrow, & Swank, 1999	Implementation data were taken from the state validation process completed at the end of a school's second year using the IST process. Phase 1 schools: the total number of components of the process that were in place was determined using a 103-item instrument. Phase 2 schools: Implementation was evaluated using an instrument that contained 7 broad areas of implementation on a 4-point scale: 0 (feature not in place), 1 (basic feature in place), 2 (feature in place at effective level), 3 (feature in place at model level). High implementation schools were the top 30% of both phase 1 and 2 schools while low implementation schools were the bottom 30%.	N/A	Students in high implementation schools had more gains in task comprehension than students in low and non schools with no significant differences between students in low and non schools. Students in high schools increased in task completion from posttest to follow-up. Students in low and non schools declined. Students in low schools had lower time on-task than students in non schools. Students in high schools made more gains than students in low and non schools from post to follow-up. Low implementation did not produce better results than none at all. Over time, students in high schools started to look like average peers on comprehension, task completion and time on-task.

Table A1 (continued)

Zvoch, Letourneau, & Parker, 2007	School district personnel using background knowledge, program manuals, and expert consensus dialogue created a 6-item checklist. Graduate students and retired educators conducted three observations during 2-week windows in October, January, and April. Scores for each teacher were averaged across the three observations to create an overall score.	N/A	A multilevel analysis was conducted. Student outcomes were assessed using a composite of DIBELS scores. Provider characteristics and class size were not significantly related to fidelity. Site-to-site differences accounted for 43 percent of the variability in students' literacy growth, and student characteristics accounted for 9 percent of the variance in students' literacy growth and 3 percent of the variance in site's initial scores and growth. The only variable that predicted differences in student outcomes was the school schedule. Fidelity scores were not related to site-based student growth rates. However, upon further analysis three low implementation high student outcomes sites were removed, and when completing the analysis after removing these three sites, a significant association between fidelity and student outcomes emerged.

Table A1 (continued)

Gersten, Carnine, & Williams, 1982	Direct observations were conducted in classrooms 4 to 6 times from November to May by trained observers using the direct instruction supervision code (DISC). Behaviors in the DISC: Accuracy of formats, time allocations, use of hand signals, pacing of lessons, student accuracy rate, and reinforcement.	As this was a validity study for the DISC, teachers were also given separate global ratings on a one to four scale.	Teachers with the highest DISC scores had low average CTBS (student outcome) scores above the national median. Teachers with the lowest DISC scores had low CTBS scores.

APPENDIX B

FIDELITY OBSERVATION FORMS

OPTIMIZE
FIDELITY OF IMPLEMENTATION CHECKLIST – SPELLING

School :_________________ Instructor:_________________ Observer ____________
Lesson #:_________________ Time began:__________
Number of children in group today: __________ Time ended:__________

Activity 1: Writer's Warm-Up______________________________		
Scoring	Critical Instructional Features	Comments
yes no partial	Used wording from script.	
yes no partial	Completed all steps in the activity.	
yes no partial	Teacher modeled new letter.	
yes no partial	All students participated with group and written responses.	
yes no partial	Teacher corrected student mistakes.	
yes no partial	Activity completed in 2 to 4 minutes.	
	Total time to complete activity:	

Activity 2: Phonologic/Alphabetic ______________________________		
Scoring	Critical Instructional Features	Comments
yes no partial	Used wording from script.	
yes no partial	Completed all steps in the activity.	
yes no partial	All students participated with group and written responses.	
yes no partial	Teacher corrected student mistakes.	
yes no partial	Teacher modeled 1 example.	
yes no partial	Teacher leads/tests students on remaining examples.	
yes no partial	Activity completed in 2 to 4 minutes.	
	Total time to complete activity:	

Activity 3: Phonologic/Spelling ______________________________		
Scoring	Critical Instructional Features	Comments
yes no partial	Used wording from script.	
yes no partial	Completed all steps in the activity.	
yes no partial	All students participated with group and written responses.	
yes no partial	Teacher corrected student mistakes.	
yes no partial	Teacher modeled 1 example.	
yes no partial	Teacher leads/tests students on remaining examples.	
yes no partial	Activity completed in 6 to 8 minutes.	
	Total time to complete activity:	

TO SCORE FIDELITY OF IMPLEMENTATION:

yes = 2 point
partial = 1 point
no = 0 points

Add all points: $\frac{\_\_\_\_\_\_\_\_\_\_\_\_}{40}$ =

GENERAL CONSIDERATIONS:		
______	Quality of Lesson Delivery (high, medium, low).	COMMENTS
______	Student Engagement (high, medium, low).	
______	Completed All Activities in the Lesson.	
______	Completed All Activities Within 15 minutes.	

Optimize-Storybook Intervention
Fidelity of Implementation Checklist (Lessons 5-6)

School____________ Instructor________________ Observer________Date______
Series and Lesson #____________________ Time began:__________
Book(s)______________________________ Time ended __________
Number of children in group today_______________

Vocabulary activities:		Comments:
yes no partial	Used wording similar to script.	
yes no partial	Lesson 5 only: Pointed effortlessly to correct illustration while reviewing vocabulary. (Part A)	
yes no partial	Lesson 6 only: Group responded during vocabulary review.	
yes no partial	Lessons 5 & 6: Group responded during games.	
yes no partial	Used correction procedures as needed.	
--------------------	Time to complete vocabulary activities.	
The Retell:		
yes no partial	Reintroduced title, author, illustrator.	
yes no partial	Children took turns in retell. Every child had an opportunity.	
yes no partial	Used designated prompts for retell (pictures/ verbal/ no prompt – see specific lesson).	
yes no partial	Corrected vocabulary use during retell.	
yes no partial	Asked all post-discussion questions.	
yes no partial	Gave each child similar opportunity to talk.	
--------------------	Time to complete "The Retell".	
General Considerations:		
yes no partial	Within 1 day of scheduled lesson.	
yes no partial	Completed entire lesson.	
high med low	Quality of lesson delivery (not scored)	
high med low	Student engagement (not scored)	

Scoring: Add 2 points if total time = 15 minutes
Yes = 2 points
Partial = 1 point Add all points: _____________ =
No = 0 points 28

OPEN COURT FIDELITY OF IMPLEMENTATION CHECKLIST

Teacher:	School:
Observer:	Date:
Lesson:	Number of Students:

Directions:

List each of the headings in red from today's lesson on the lines that follow. Then, circle yes, no, or partial to indicate if the teacher completed each step of the activity. Usually, there will be a red bullet to indicate each step. Please note that sometimes there are steps that are not indicated by bullets. This is rare, but usually happens at the very beginning or the very end of an activity. Please do include these non-bulleted directions in your checklist.

1. ________________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

2. ________________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

3. ________________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

4. ______________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

5. ______________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

6. ______________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

7. ______________________________ TIME: ____________________

COMMENTS

- yes no partial
- yes no partial
- yes no partial
- yes no partial
- yes no partial

READ ALOUD COMPONENT:
(Applicable for books B-E.)

Directions:
List the activities from Marsha's fax that correspond to today's lesson. Then, indicate whether the teacher completed the activity by circling yes, no, or partial. Document the number of minutes spent on the read aloud component.

1. ______________________________ yes no partial

2. ______________________________ yes no partial

3. ______________________________ yes no partial

COMMENTS:

Time spent on READ ALOUD activities (in minutes): ____________
(Should be about 5 minutes.)

TOTAL TIME ON LESSON (in minutes) = ___/30

TO SCORE FIDELITY OF IMPLEMENTATION:

yes = 2 points partial = 1 point no = 0 points

Add all points, then divide by total possible to determine fidelity of implementation.

$\frac{\text{points earned}}{\text{total possible points}}$ = fidelity

GENERAL CONSIDERATIONS:

__________ **Quality of Lesson Delivery (high, medium, low)**

__________ **Student Engagement (high, medium, low)**

__________ **Completed All Activities In the Lesson.**

__________ **Completed all Activities Within 30 Minutes.**

COMMENTS:

APPENDIX C

RESULTS TABLES

Table C1

Descriptive Statistics for Fidelity Measures at the Individual Observation Level

	M	*SD*	*Skew*	*Kurt*	*Mdn*	*Min*	*Max*	*N*
Total Fidelity	90.14	9.02	-1.70	4.83	91.64	44.17	100.00	187
Total Infidelity	2.71	1.59	-0.10	-0.20	2.89	0.00	7.47	187
Quality	2.73	0.44	-1.49	1.74	3.00	1.00	3.00	142
Trimmed Quality	2.74	0.40	-1.06	-0.60	3.00	2.00	3.00	142
Engagement	2.76	0.38	-1.21	-0.24	3.00	2.00	3.00	142

Table C2
Engagement Growth Model Results

	Model 1	Model 2	Model 3
Effect	estimate	estimate	estimate
Intercept	2.749***	2.748***	2.748***
Slope	0.004	0.004	0.004
SD(Intercept)	0.250***	0.196***	0.196***
SD(Slope)	0.024	0.000	
cor(Intercept, Slope)	-1.000***		
SD(residual)	0.335***	0.338***	0.338***
loglikelihood	-64.005	-64.754	-64.754

Model Comparisons	1 vs 2	1 vs 3
χ^2	1.498	1.498
df	1	2
p	0.221	0.473

Model Comparisons	2 vs 3
χ^2	0.000
df	1
p	1.000

***$p < .001$.

Table C3

Quality Growth Model Results

Effect	Model 1 estimate	Model 2 estimate	Model 3 estimate
Intercept	2.654***	2.646***	2.646***
Slope	0.039*	0.042**	0.042**
SD(Intercept)	0.362***	0.269***	0.269***
SD(Slope)	0.041	0.000	
cor(Intercept, Slope)	-0.904***		
SD(residual)	0.286***	0.298***	0.298***
loglikelihood	-52.325	-55.050	-55.050

Model Comparisons		1 vs 2	1 vs 3
χ^2		5.450	5.450
df		1	2
p		0.020	0.066

Model Comparisons			2 vs 3
χ^2			0.000
df			1
p			1.000

* $p < .05$.
** $p < .01$.
***$p < .001$.

Table C4

Infidelity Growth Model Results

Effect	Model 1 estimate	Model 2 estimate	Model 3 estimate
Intercept	3.102***	3.103***	3.101***
Slope	-0.189**	-0.189**	-0.188***
SD(Intercept)	0.789***	0.772***	0.810***
SD(Slope)	0.135	0.128	
Cor(Intercept, Slope)	-0.081		
SD(residual)	1.318***	1.320***	1.344***
loglikelihood	-338.831	-338.838	-339.255

Model Comparisons	1 vs 2	1 vs 3
χ^2	0.013	0.848
df	1	2
p	0.910	0.655

Model Comparisons	2 vs 3
χ^2	0.835
df	1
p	0.361

** *p < .01.*

****p* < .001

Table C5

Multi-Process, Multi-Level Fidelity Model

	Model 1	Model 2	Model 3	Model 4
Effect	estimate	estimate	estimate	estimate
Infidelity	3.117***	3.110***	3.117***	3.101***
Quality	2.654***	2.653***	2.657***	2.646***
Engagement	2.753***	2.747***	2.748***	2.748***
Infidelity Slope	-0.189***	-0.187***	-0.189***	-0.188***
Quality Slope	0.042**	0.042**	0.042**	0.042**
Engagement Slope	0.004	0.004	0.004	0.004
SD Infidelity	0.824***	0.822***	0.824***	0.810***
SD Quality	0.266***	0.249***	0.268***	0.269***
SD Engagement	0.189***	0.193***	0.196***	0.196***
cor Infidelity-Quality	-0.798***	-0.719***	-0.801***	
cor Quality-Engagement	0.689**	0.462		
cor Infidelity-Engagement	-0.486			
SD Quality Residual	0.298***	0.298***	0.297***	0.298***
SD Engagement Residual	0.340***	0.339***	0.338***	0.338***
SD Infidelity Residual	1.342***	1.343***	1.342***	1.344***
loglikelihood	-450.103	-451.509	-453.101	-459.059
Model Comparisons		1 vs 2	1 vs 3	1 vs 4
χ^2		2.811	5.995	17.911
df		1.000	2.000	3.000
p		0.094	0.050	0.000
Model Comparisons			2 vs 3	2 vs 4
χ^2			3.184	15.100
df			1.000	2.000
p			0.074	0.001
Model Comparisons				3 vs 4
χ^2				11.916
df				1.000
p				0.001

Table C6

Between and Within Group Variance Components for May Outcomes

					Variance	Nested		
	Est.	*SE*	Est./*SE*	*p*	Proportion	χ^2	*df*	*p*
Means								
ORF 1	2.49***	0.25	10.14	0.00				
ORF 2	2.04***	0.18	11.35	0.00				
NWF	5.34***	0.18	29.56	0.00				
PSF	41.46***	2.15	19.32	0.00				
Word Attack	107.38***	1.38	78.08	0.00				
Word ID	102.50***	1.39	73.62	0.00				
Within Group Variance								
ORF 1	2.11***	0.37	5.70	0.00	0.68			
ORF 2	1.94***	0.36	5.43	0.00	0.87			
NWF	2.31***	0.41	5.67	0.00	0.91			
PSF	232.19***	39.22	5.92	0.00	0.81			
Word Attack	111.05***1	18.94	5.86	0.00	0.86			
Word ID	124.90***	21.43	5.83	0.00	0.89			
Between Group Variance								
ORF 1	0.98*	0.50	1.97	0.05	0.32	7.10	1	0.01
ORF 2	0.29	0.32	0.90	0.37	0.13	0.93	1	0.33
NWF	0.22	0.32	0.69	0.49	0.09	0.55	1	0.46
PSF	55.40	36.73	1.51	0.13	0.19	3.86	1	0.05
Word Attack	17.49	15.17	1.15	0.25	0.14	1.89	1	0.17
Word ID	15.00	16.40	0.91	0.36	0.11	0.99	1	0.32

* $p < .05.$

***$p < .00$

Table C7

Infidelity Predicting Group Level Student Outcome in May

	Outcome					
Effect	Word ID	Word Attack	NWF	PSF	ORF passage 2	ORF passage 1
Outcome Within Group Variance	122.92***	109.91***	2.26***	228.67***	1.91***	2.08***
Infidelity Residual Variance	1.43***	1.43***	1.43***	1.43***	1.43***	1.43***
Infidelity Group Variance	0.64*	0.64*	0.64*	0.64*	0.64*	0.64*
Spring Infidelity Intercepts	-0.66**	-0.66**	-0.66**	-0.66**	-0.66**	-0.66**
Infidelity Mean	0.15	0.15	0.15	0.15	0.15	0.15
Outcome Intercept	101.98***	106.71***	5.27***	40.68***	1.97***	2.42***
Outcome Group Residual Variance	8.13	3.53	0.13	40.02	0.14	0.86
Infidelity effect on Outcome	3.83	4.91*	0.51	5.87	0.55	0.53
Proportion Group Variance	0.12	0.15	0.11	0.21	0.15	0.33
Group R^2	0.54	0.81	0.56	0.35	0.59	0.17

* *p < .05.*

** *p < .01.*

****p* < .001.

Table C8

Engagement Predicting Group Level Student Outcome in May

	Outcome					
Effect	Word ID	Word Attack	NWF	PSF	ORF passage 2	ORF passage 1
Outcome Within Group Variance	124.09***	109.25***	2.30***	230.15***	1.91***	2.08***
Engagement Residual Variance	0.07***	0.07***	0.07***	0.07***	0.07***	0.07***
Engagement Mean Group Variance	0.05*	0.05*	0.05*	0.05*	0.05*	0.05*
Engagement Mean	0.05	0.05	0.05	0.05	0.05	0.05
Outcome Intercept	101.95***	107.01***	5.27***	40.58***	1.96***	2.37***
Outcome Group Residual Variance	8.64	16.19	0.07	37.55	0.14	0.64
Engagement Mean effect on Outcome	12.17	9.05	1.80	21.10	1.96*	2.79*
Proportion Group Variance	0.11	0.16	0.09	0.21	0.15	0.33
Group R^2	0.46	0.20	0.69	0.37	0.59	0.38

* $p < .05$.

***$p < .001$.

Table C9

Quality (Average Level and Slope) Predicting Group Level Student Outcome in May

	Outcome					
Effect	Word ID	Word Attack	NWF	PSF	ORF passage 2	ORF passage 1
Outcome Within Group Variance	121.25***	110.41***	2.28***	271.78***	1.93***	2.10***
Quality Residual Variance	0.07***	0.07***	0.07***	0.09***	0.07***	0.08***
Quality Mean Group Variance	0.10**	0.11**	0.11**	0.12**	0.11**	0.11**
Quality Mean-Slope Group Covariance	-0.21*	-0.23*	-0.22*	-0.27**	-0.22*	-0.23*
Quality Mean-Slope Group Correlation	-0.91*	-0.97*	-0.97*	-1.29**	-0.97*	-0.99*
Quality Slope Group Variance	0.53	0.50	0.49	0.35	0.49	0.48
Quality Mean	0.00	0.01	0.00	0.01	0.00	0.00
Quality Slope	0.45*	0.44*	0.45*	0.46*	0.46*	0.48*
Outcome Intercept	96.10***	97.64***	3.92	41.08***	0.49	-1.84
Outcome Group Residual Variance[a]	0	0	0	0	0	0
Quality Mean effect on Outcome	25.87	39.05	6.03	-8.88	6.48	18.49
Quality Slope effect on Outcome	14.14	21.47	3.08	0.97	3.36	9.13
Proportion Group Variance	0.13	0.14	0.10	0.05	0.14	0.32
Group R^2	1.00	1.00	1.00	1.00	1.00	1.00

Notes: [a]Fixed to zero to prevent convergence problems.

* $p < .05$.

** $p < .01$.

***$p < .001$.

Table C10

Quality (Initial Status and Slope) Predicting Group Level Student Outcome in May

	Outcome					
Effect	Word ID	Word Attack	NWF	PSF	ORF passage 2	ORF passage 1
Outcome Within Group Variance	126.72***	112.56***	2.31***	234.42***	1.95***	2.11***
Quality Residual Variance	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***
Quality Initial Status Group Variance	0.28**	0.29**	0.29**	0.29**	0.29**	0.29**
Quality Initial Status	-0.14	-0.13	-0.14	-0.13	-0.14	-0.15
Quality Slope	0.59**	0.58**	0.59**	0.58**	0.59**	0.60**
Quality Initial Status effect on Slope	-2.55***	-2.56***	-2.54***	-2.54***	-2.55***	-2.54***
Outcome Intercept	101.75***	106.05***	5.26***	39.22***	1.93***	2.42***
Outcome Group Residual Variance	11.36	11.76	0.21	41.68	0.25	0.96
Quality Slope effect on Outcome	0.70	1.36	0.08	2.36	0.11	0.07
Proportion Group Variance	0.09	0.12	0.09	0.18	0.12	0.32
Group R^2	0.07	0.23	0.05	0.20	0.09	0.01

** *p < .01.*

***$p < .001$.

APPENDIX D

RESULTS FIGURES

Figure D1. Trimmed and transformed fidelity constructs.

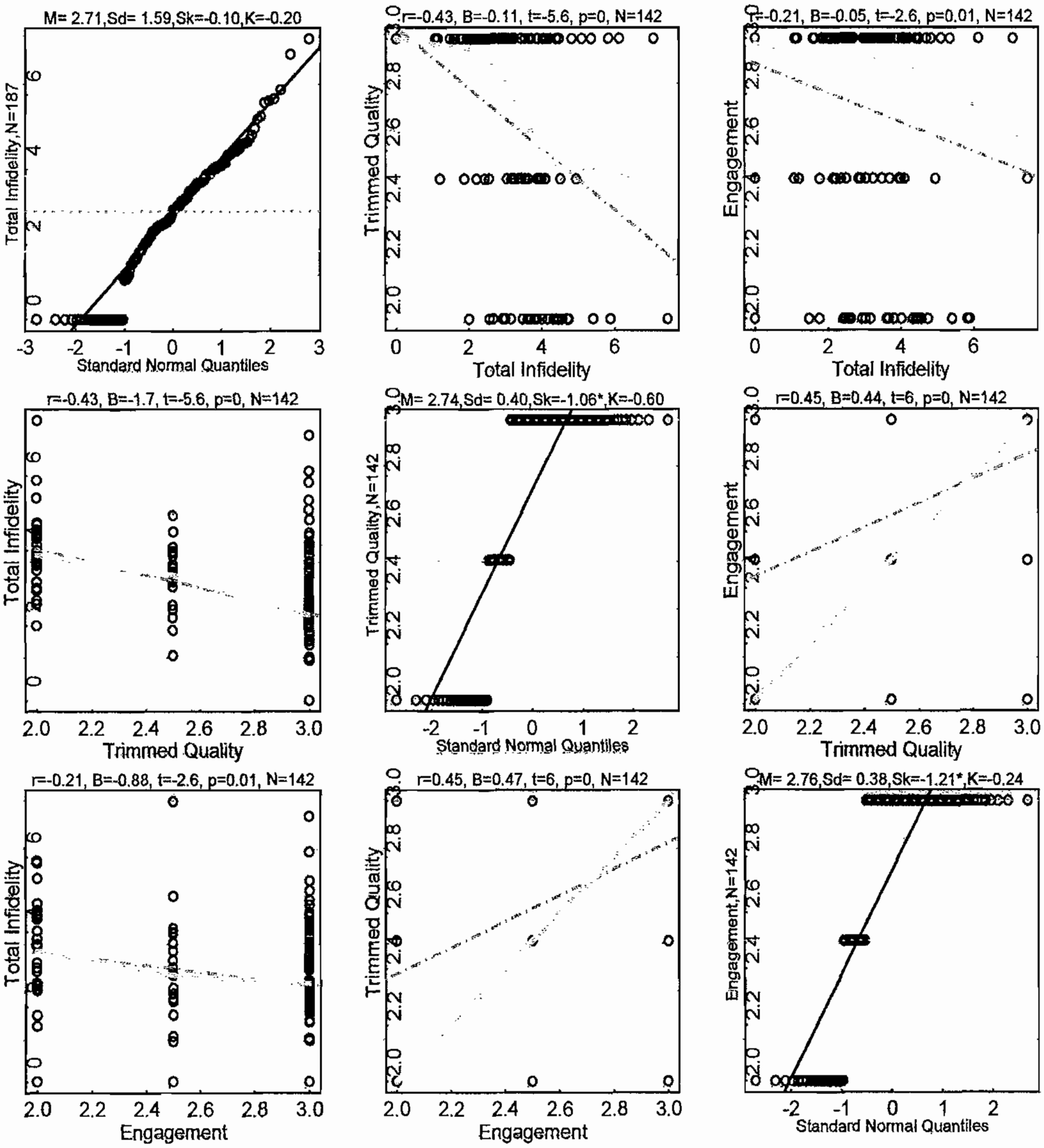


Figure D2. Total infidelity growth curves for each group.

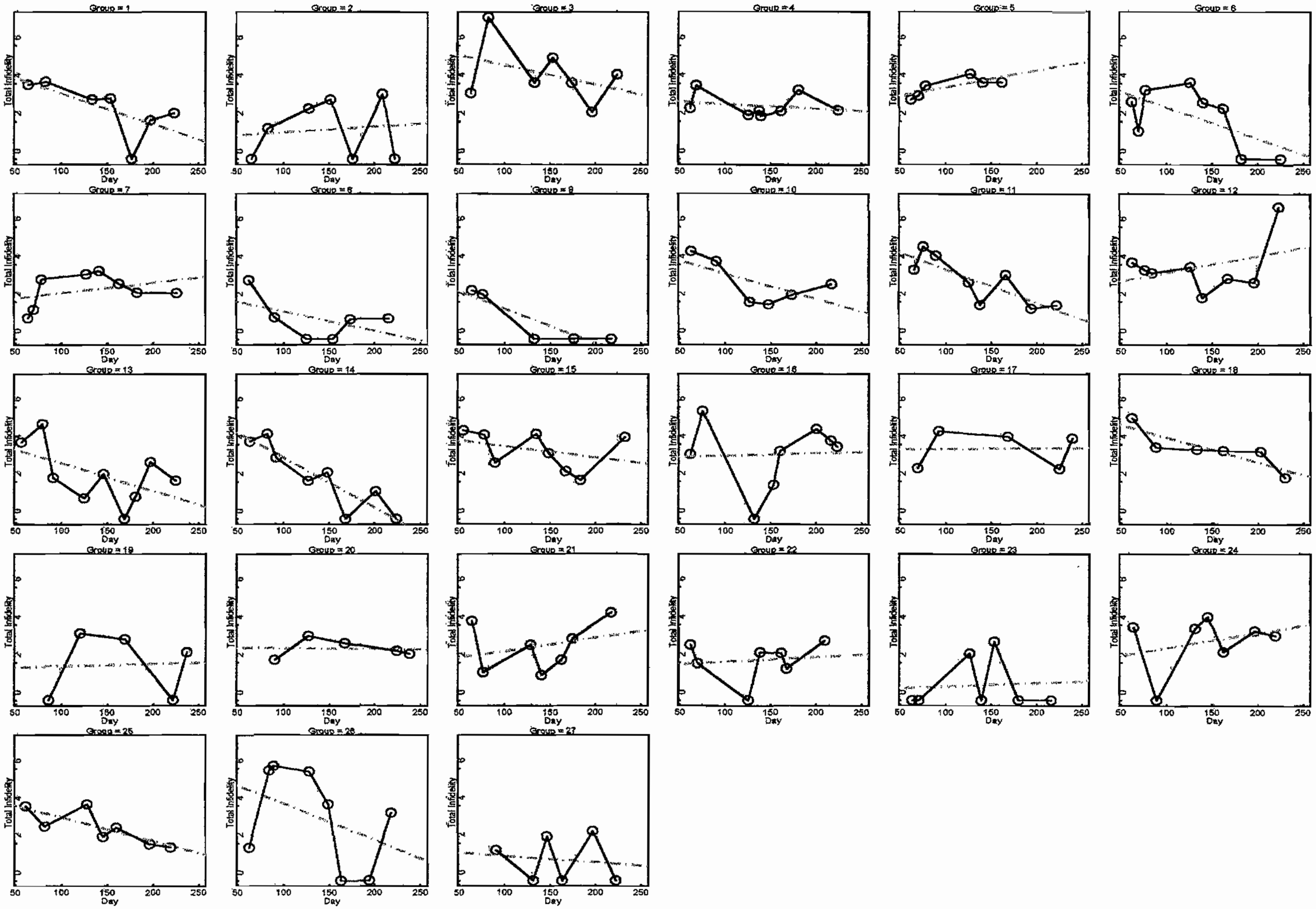


Figure D3. Trimmed quality growth curves for each group.

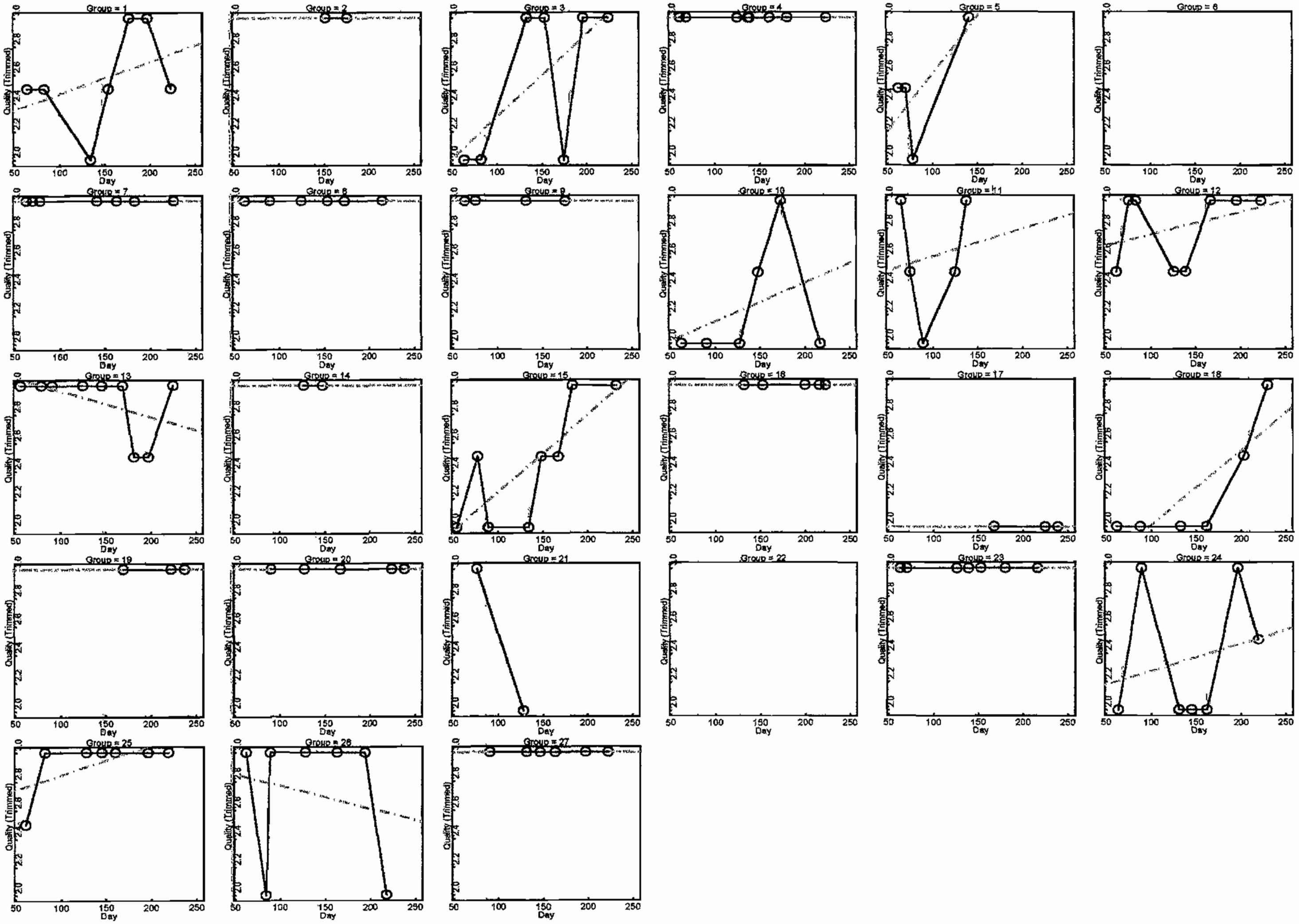
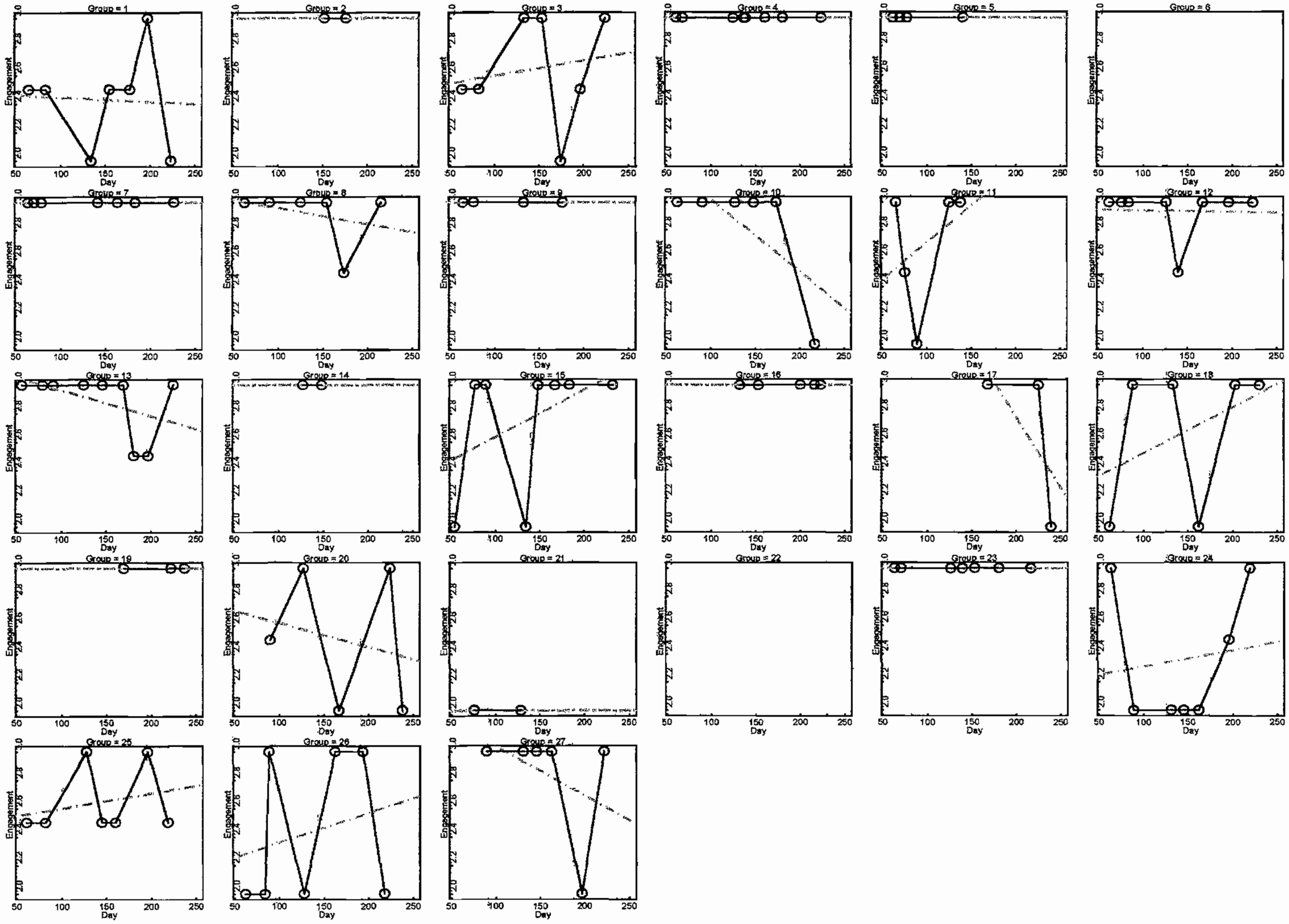


Figure D4. Engagement growth curves for each group.



BIBLIOGRAPHY

Adams, M. J., Bereiter, C., Brown, A., Campione, J., Caruthers, I., Case, R., et al. (2000). *Open Court Reading.* New York, NY: McGraw-Hill.

Alberto, P. A. & Troutman, A. C. (1995). *Applied behavior analysis for teachers* (6th ed.). Englewood cliffs, NJ: Merrill/Prentice Hall.

Al Otaiba, S., & Fuchs, D. (2006). Who are the young children for whom best practices in reading are ineffective?: An experimental and longitudinal study. *Journal of Learning Disabilities, 39*, 414-431.

Batsche, G., Elliott, J., Graden, J. L., Grimes, J., Kovaleski, J. F., Prasse, D., Reschly, D. J., Schrag, J., & Tilly, W. D. (2006). *Response to intervention policy considerations and implementation.* Alexandria, VA: National Association of State Directors of Special Education, Inc.

Bellg, A. J., Borelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., et al. (2004). Enhancing treatment fidelity in heath behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology, 23*(5), 443-451.

Borman, G. D., Hewes, G., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73(2),* 125-230.

Brophy, J. E. & Good, T. L. (1986). Teacher behavior and student achievement. In M. D. Wittrock (Ed.), *Handbook of research on teaching, 3rd ed.,* (pp. 328-375). New York, NY; Macmillan.

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2004). *Direct Instruction Reading* (4th ed.). Upper Saddle River, NJ: Pearson.

Castro, F. G., Barrera, M., Jr., & Martinez, C. R., Jr. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science, 5*(1), 41-45.

Chard, D. J., & Harn, B. A. (in press). Project CIRCUITS: Center for Improving Reading Competence Using Intensive Treatments Schoolwide.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis in the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Denton, C. A., Vaughn, S., & Fletcher, J. M. (2003). Bringing research-based practice in reading intervention to scale. *Learning Disabilities Research & Practice, 18*, 201-211.

Dobson, K. S. & Singer, A. R. (2005). Definitional and practical issues in the assessment of treatment integrity. *Clinical Psychology: Science and Practice, 12*(4), 384-387.

Dunn, L. & Dunn, L. (1981). *Peabody picture vocabulary test-Revised.* Circle Pines, MN: American Guidance Service.

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

Gersten, R. M., Carnine, D. W., & Williams, P. B. (1982). Measuring implementation of a structured educational model in an urban school district: An observational approach. *Educational Evaluation and Policy Analysis*, *4,* 67-79.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71,* 149-164.

Gettinger, M., & Stoiber, K. C. (2006). Functional assessment, collaboration, and evidence-based treatment: Analysis of a team approach for addressing challenging behaviors in young children. *Journal of School Psychology, 44*, 231-252.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV* (pp. 699-720). Bethesda, MD: National Association of School Psychologists.

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

Good, R. H., & Kaminski, R. A. (2003). *DIBELS™: Dynamic Indicators of Basic Early Literacy Skills, Sixth Edition.* Longmont CO: Sopris West.

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

Gresham, F. M., Gansle, K., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis, 26,* 257-263.

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

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice, 15,* 198-205.

Haager, D., Klingner, J., & Vaughn, S. (2007). *Evidence-Based Reading Practices for Response to Intervention.* Baltimore: Paul H. Brookes Publishing Co.

Halle, J. (1998). Fidelity: A crucial question in translating research to practice. *Journal of Early Intervention, 21,* 294-296.

Harn, B. A., Chard, D. J., Kame'enui, E. J., Allen, M., & Parisi, D. (in press). School-level reading experiences: Examining the role of instructional practices on preventing long-term reading difficulties. *Learning Disabilities Research & Practice.*

Harn, B. A., Kame'enui, E. J., & Simmons, D. C. (2007). The nature and role of the third tier in a prevention model for kindergarten students. In D. Haager, J. Klingner, & S. Vaughn (Eds.), *Evidence-based reading practices for response to intervention* (pp. 161-184). Baltimore, MD: Paul H. Brookes Publishing.

Hayes, S. C., Nelson, R. O., & Jarret, R. B. (1986). Evaluating the quality of behavioral assessment. In R. Nelson & S. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 461-503). New York: Guilford.

Howell, K., & Nolet, V. (2000). *Curriculum-based evaluation: Teaching and decision making* (3rd ed.). Australia: Wadsworth.

Individuals with Disabilities Act Regulations, 34 C.F.R. 300 (2004)

Jones, K. M., Wickstrom, K. F., & Friman, P. C. (1997). The effects of observational feedback on treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 12,* 316-326.

Klingner, J. K. (2004). The science of professional development. *Journal of Learning Disabilities, 37*, 248-255.

Kovaleski, J. F., Gickling, E. E., Morrow, H., & Swank, P. R. (1999). High versus low implementation of instructional support teams: A case for maintaining program fidelity. *Remedial and Special Education, 20*, 170-183.

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. *Preventing School Failure, 48,* 36-43.

LeLaurin, K., & Wolery, M. (1992). Research standards in early intervention: Defining, describing, and measuring the independent variable. *Journal of Early Intervention, 16,* 275-287.

Leventhal, H. & Friedman, M. A. (2004). Does establishing fidelity of treatment help in understanding treatment efficacy? Comment on Bellg et al. (2004). *Health Psychology, 23*(5), 452-456.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11,* 247-266.

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

Muthén, L. K. and Muthén, B. O. (2007). Mplus. (Fifth Edition.) [Computer software.] Los Angeles, CA: Muthén & Muthén.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* Bethesda, MD: National Institute of Child Health and Human Development.

National Joint Committee on Learning Disabilities. (2005, June). *Responsiveness to interventions and learning disabilities.*

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*(1), 33-84.

Orwin, R.G. (2000). Assessing program fidelity in substance abuse health services research. *Addiction, 95*(3), 309-327.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*, 365-383.

Persampieri, M., Gortmaker, V., Daly, E. J., Sheridan, S. M., & McCurdy, M. (2006). Promoting parent use of empirically supported reading interventions: Two experimental investigations of child outcomes. *Behavioral Interventions, 21*, 31-57.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*, 477-492.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus.* New York, NY: Springer Publishing Co.

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

R Development Team. (2007). The R Project for Statistical Computing (Version 2.6.0) [Computer software]. Vienna, Austria: R Development Team.

S-Plus. (2006). S-Plus (Version 8) [Computer software]. Seattle, WA: Insightful.

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9th Ed.). Boston, MA: Houghton Mifflin Co.

Shinn, M. R. (1989). *Curriculum-Based Measurement: Assessing Special Children.* New York, NY: Guilford Press.

Simmons, D. C., Kame'enui, E. J., Harn, B., Coyne, M. D., Stoolmiller, M., Santoro, L. E., et al. (2007). Attributes of effective and efficient kindergarten intervention: An examination of instructional time and design specificity. *Journal of Learning Disabilities, 40*(4), 331-347.

Smith, S. W., Daunic, A. P., & Taylor, G. G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children, 30*(4), 121-134.

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 Consulting and Clinical Psychology, 68*(2) 296-306.

Telzrow, C. F., McNamara, K., & Hollinger, C. L. (2000). Fidelity of problem-solving implementation and relationship to student performance. *School Psychology Review, 29*, 443-461.

Tindal, G., Marston, D., & Deno, S. (1983). The reliability of direct and repeated measurement. *Journal of Special Education Leadership, 12*(2), 3–10.

Torgesen, J. K. (2001). The theory and practice of intervention: Comparing outcomes from prevention and remediation studies. In A. J. Fawcett (Ed.), *Dyslexia: Theory and good practice.* London: Whurr Publishers.

Vaughn, S., & Dammann, J. E. (2001). Science and sanity in special education. *Behavioral Disorders 27*(1), 21-29.

Vaughn, S., & Fuchs, L. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137-146,

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69*, 391-409.

van Otterloo, S. G., van der Leij, A., & Veldkamp, E. (2006). Treatment integrity in a home-based pre-reading intervention programme. *Dyslexia, 12*, 155-176.

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

Witt, J. C., Noell, G. H., LaFleur, L. H., & Mortenson, B. P. (1997). Teacher use of interventions in general education settings: Measurement and analysis of the independent variable. *Journal of Applied Behavior Analysis, 3,* 696-696.

Woodcock, R. (1987). *Woodcock reading mastery test-Revised.* Circle Pines, MN: American Guidance Service.

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

Ysseldyke, J., & Christenson, S.L. (1989). Linking assessment to intervention. In J. Graden, J. Zins, M. Curtis (Eds.), *Alternative educational delivery systems* (p. 91-110). Bethesda, MD: NASP Publications

Zvoch, K., Letourneau, L. E., & Parker, R. P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation, 28*, 132-150.