MEASURING INSTRUCTIONAL INTERACTIONS IN KINDERGARTEN MATHEMATICS CLASSROOMS THROUGH A DIRECT OBSERVATION SYSTEM

by

CHRISTIAN T. DOABLER

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

March 2010

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

Christian Doabler

Title:

"Measuring Instructional Interactions in Kindergarten Mathematics Classrooms Through a Direct Observation System"

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

Robert Horner, Chairperson, Special Education and Clinical Sciences
Christopher Murray, Member, Special Education and Clinical Sciences
Scott Baker, Member, Special Education and Clinical Sciences
Joe Stevens, Member, Educational Methodology, Policy, and Leadership
Thomas Dishion, Outside Member, Psychology

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

March 20, 2010

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

An Abstract of the Dissertation of

Christian T. Doabler for the degree of Doctor of Philosophy

in the Department of Special Education and Clinical Sciences

to be taken March 2010

Title: MEASURING INSTRUCTIONAL INTERACTIONS IN KINDERGARTEN MATHEMATICS CLASSROOMS THROUGH A DIRECT OBSERVATION SYSTEM

Approved: ______________________________

Dr. Robert Horner, Chair

There is convincing evidence that many students struggle to learn mathematics proficiently. One plausible contributor to the low math achievement is the quantity and quality of learning opportunities provided in classrooms. These opportunities may fall short of addressing the learning needs of students, especially those at risk for failure in mathematics.

Against this backdrop, the purpose of the dissertation was to validate a direct observation instrument. The Coding of Academic Teacher-Student interactions (CATS) observation instrument systematically measures the instructional interactions that occur between teachers and students during kindergarten mathematics instruction. The dissertation harvested data from the Early Learning in Mathematics: Efficacy Trials in

Kindergarten Classrooms (ELM-ETKC) project, a randomized control efficacy trial. ELM-ETKC is investigating the efficacy of the Early Learning in Mathematics curriculum within 65 kindergarten classrooms across three school districts in the state of Oregon.

The dissertation utilized student and classroom-level information collected in 65 ELM-ETKC kindergarten classrooms across the 2008-2009 school year. At the student level, data included scores from 929 kindergarten students on the Test of Early Mathematics Ability-Third Edition (TEMA) and two curriculum-based measures: Oral Counting and Number Identification. Information at the classroom level included observational data from 191 classroom observations.

Utilizing the extant data, the dissertation addressed research questions related to content validity, discriminant validity, and criterion-predictive validity. Additionally, the study examined if observers could reliably use the CATS instrument in classrooms. To address two of the research questions, the dissertation employed a hierarchical design and fit multilevel models that nested (a) observations within classrooms and (b) student posttest TEMA scores within classrooms. Predictors of the models included student risk status and rates of observed instructional behaviors.

The study found promising evidence for using the CATS instrument to collect information about the quantity and quality of kindergarten mathematics instruction. Independent observers reached acceptable interobserver agreement across the observations. The CATS instrument demonstrated high levels of content validity, as well as sensitivity to treatment conditions. Results also found statistically significant relationships between the mean rate of instructional behaviors and student posttest TEMA scores. Implications for future research and practice are provided.

CURRICULUM VITAE

NAME OF AUTHOR: Christian T. Doabler

PLACE OF BIRTH: Vineland, New Jersey

DATE OF BIRTH: August 22, 1971

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Rowan University, Glassboro, NJ

DEGREES AWARDED:

Doctor of Philosophy in Special Education and Clinical Sciences, 2010, University of Oregon
Master of Science in Special Education, 2003, University of Oregon
Bachelor of Arts, 1995, Rowan University

AREAS OF SPECIAL INTEREST:

Prevention of Learning Disabilities
Reading and Mathematics Instruction
Design of Instruction

PROFESSIONAL EXPERIENCE:

Research Assistant, Pacific Institutes for Research, Eugene, Oregon, 2009-present

Graduate Fellow, Center on Teaching and Learning, Eugene, Oregon, 2008-2009

Teaching Assistant, Department of Special Education and Clinical Services, University of Oregon, Eugene, 2006-2007

Special Educator, Springfield Public Schools, Springfield, Oregon, 2003-2006

PUBLICATIONS:

Doabler, C. T., Fien, H., Nelson-Walker, N., & Baker, S. (2010). *Evaluating the instructional design elements of elementary mathematics programs.* Manuscript submitted for publication.

Baker, S. K., Chard, D. J., Ketterlin-Geller, L., Apichatabutra, C., & Doabler, C. T. (2009). Teaching writing to at-risk students: The quality of evidence for self-regulated strategy development. *Exceptional Children, 75*, 303–320.

Chard, D. J., Ketterlin-Geller, L., Baker, S., Doabler, C. T., & Apichatabutra, C. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children, 75*, 263–284.

Doabler C. T. (2008). What CEC students should know about effective reading instruction. *Inspire and Inquire—Council for Exceptional Children Newsletter 2*(2), 4–5.

ACKNOWLEDGMENTS

To my friends and family, especially, my mom, Roy, Barry, and Jaime, thank you for your enduring support. I'm also grateful to my good friend, Hank, whose judicious feedback and words of encouragement allowed me to finish this long, strange trip. Special thanks are also due to Drs. Scott Baker and Keith Smolkowski for sharing their time and invaluable expertise in educational research. Finally, and above all, I want to thank my loving wife, Krista, for always believing in me.

For Francis J. Fien III, a great friend.
“When they built you, brother, they broke the mold”

TABLE OF CONTENTS

Chapter Page

I. STATEMENT OF THE PROBLEM 1

A Model of School Learning 3

Measurement of Teaching 6

Purpose of the Study 9

Research Questions 10

II. A REVIEW OF THE LITERATURE 13

Practice Opportunities 13

Explicit Instruction 20

Summary and Linkages to Present Study 24

III. METHODOLOGY 26

Design 27

The ELM-ETKC Project 27

Participants 28

Dissertation Sample 30

Coding of Academic Teacher-Student Interactions (CATS) Observation Instrument 32

Observer Training and Interobserver Agreement 38

Kindergarten Mathematics Instruction 41

Chapter Page

Measures 42

Data Analysis 48

IV. RESULTS 58

Missing Student Data 58

Descriptive Statistics 59

Diagnostics of the Multilevel Models 63

Intercorrelations 63

Content Validity 64

Interobserver Agreement 66

Sensitivity to Detect Differences Between Treatment Conditions 69

Prediction of Student Outcomes 75

V. DISCUSSION 85

Summary of Results 86

Evidence of Content Validity 86

Consistency of the Observation Data 88

Discriminant Validity 90

Prediction of Student Outcomes 92

Limitations 94

Implications for Practice 97

Implications for Research 98

Conclusion 99

Chapter Page

APPENDIX: CODING OF ACADEMIC TEACHER-STUDENT INTERACTIONS DIRECT OBSERVATION INSTRUMENT 100

REFERENCES 102

LIST OF TABLES

Table Page

1. Carroll's (1963) Model of School Learning.......... 4

2. Descriptive Statistics Per Observation Round for Average Rate of Observed Behaviors Across Treatment Conditions 60

3. Descriptive Statistics for Average Rate of Observed Behaviors Aggregated Across Observation Rounds.......... 61

4. Descriptive Statistics of Student Pretest and Posttest Performances on Test of Early Mathematics Ability-Third Edition (TEMA), and Pretest Performances on Oral Counting (OC) and Number Identification (NI) Curriculum-Based Measures 62

5. Descriptive Statistics for Context Codes Across Conditions Per Observation Round 62

6. Intercorrelations Among Rate of Observed Behaviors.......... 64

7. Interobserver Agreements for 46 Paired (Reliability) Observations.......... 67

8. Fixed Effects From Unconditional Model for Rate of Observed Behaviors 70

9. Variance Components From Unconditional Models for Rates of Observed Behaviors.......... 71

10. Fixed Effects from Conditional Models for Rates of Observed Behaviors.......... 73

11. Variance Components from Conditional Models for Rates of Observed Behaviors.......... 74

12. Variance Components From Unconditional Model for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores 76

Table Page

13. Fixed Effects From Conditional Model With Level 1 Predictors for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scors............................ 78

14. Variance Components From Conditional Model With Level 1 Predictors for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores............... 78

15. Fixed Effects From the Intercepts- and Slopes-As-Outcomes Model for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores....................... 80

16. Variance Components From the Intercepts- and Slopes-as-Outcomes Models for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores......... 81

17. Summary Table of Support for Coding of Academic Teacher-Student Interactions (CATS) Observation Instrument... 87

LIST OF FIGURES

Figure Page

1. An Example of an Instructional Interaction Between a Teacher and Their Students 16

2. A Coding Example of an Instructional Interaction 36

3. Predicted Posttest TEMA Scores for Students With and Without Math Difficulties in Four Different Types of Classrooms 83

4. Cover Sheet for the Coding of Academic Teacher-Student Interactions Instrument 100

5. Instructional Interaction Codes Captured by the Coding of Academic Teacher-Student Interactions Instrument 101

CHAPTER I

STATEMENT OF THE PROBLEM

To keep pace with the productivity of today's global marketplace, graduating high school students must exit with a deep understanding of fundamental mathematics (Committee on Prospering in the Global Economy of the 21st Century, 2007; Freidman, 2009; Glen Commission, 2000; Goldin & Katz, 2008; Gonzales et al., 2004; Levy & Murnane, 2004; Ma, 1999; National Mathematics Advisory Panel [NMAP], 2008). Students require not only a formal knowledge of whole numbers, but also proficiency in key mathematical domains such as fractions, decimals, and algebra (Ketterlin-Geller, Jungjohann, Chard, & Baker, 2007; Wu, 2008). Thus, as noted by the Mathematics Learning Study Committee (National Research Council [NRC], 2001), "*All young Americans must learn to think mathematically, and they must think mathematically to learn*" (p. 16).

Because of these heightened expectations for academic success, all children deserve the opportunity to become mathematically literate (Murnane & Levy, 1996; NRC, 2001). Yet, concern over the lack of mathematics proficiency among American students has grown exponentially over the last 15 years. A considerable amount of these worries stem from the large number of students struggling to learn the fundamental areas of beginning mathematics (e.g., base-ten system). Evidence suggests that 4-8% of the school-age population exhibits some type of mathematical disability (Fuchs, Fuchs, & Prentice, 2004; Fuchs et al., 2008; Geary, 2004; Swanson & Jerman, 2006). For many of these students, difficulties involve mastery of arithmetic combinations (e.g., automatic retrieval of basic addition problems) and efficiency of counting strategies (e.g., counting-on procedure; Geary, 1993; Gersten, Jordan, & Flojo, 2005).

Equally striking is the amount of students, both with and without mathematics learning disabilities, who fail to perform at a level commensurate with their typical achieving peers. The recent performances of American students on the Trends in International Mathematics and Science Study (TIMSS) provide a troubling example. Results from the 2003 TIMSS ranked U.S. Grade 8 students 15th out of 44 nations (Ferrini-Mundy & Schmidt, 2005; Gonzales et al., 2004). Data from the 2007 National Assessment of Educational Progress (NAEP) indicate that only 38% of Grade 4 and 35% of Grade 8 students performed at or above the *Proficient* level in mathematics (National Center of Educational Statistics [NCES], 2007). On the same NAEP assessment, only 19% and 8% of students with disabilities, respectively, scored at or above proficiency in Grades 4 and 8 (NCES, 2007).

While the long-term implications of these performances are unclear, it is transparent that too many students are struggling to become proficient in the fundamentals of school mathematics. It is therefore critical that students set foot on an early pathway for learning success. There is general consensus that this pathway must begin in kindergarten (Bodovski & Farkas, 2007; Chard et al., 2008; Denton & West, 2002; Guarino, Hamilton, Lockwood, & Rathbun, 2006; Rathbun & West, 2004; West, Denton, & Germino-Hausken, 2001).

A Model of School Learning

For the past 50 years, researchers have directed much attention toward improving the quality of classroom instruction. In 1963, for example, Carroll proposed a model of school learning to address the issue of instructional quality. Though Carroll offered the model over 45 years ago, many still consider it a plausible framework for thinking about the improvement of instruction and the promotion of student achievement (e.g., Kame'enui & Simmons, 1990; Simmons & Kame'enui, 1996; Simmons et al., 1998; Simmons et al., 2007; Snow, Burns, & Griffin, 1998).

The Carroll (1963) model fosters a practical view of the variations of school learning. Central to the model is the influence of time and the opportunities to learn that students receive. Carroll conceptualized the act of learning as the amount of time required to meet learning needs. This conceptualization is best illustrated in the equation below, where the degree of learning is a function of the ratio of the amount of time the learner is engaged in learning (TE) to the amount of time the learner requires to master a given task (TR). The equation is as follows:

$$\text{Degree of learning} = f\left(\frac{TE}{TR}\right)$$

Within Carroll's (1963) model of school learning are five variables, each focusing specifically on the amount of time actually spent on learning. Table 1 illustrates these variables and their corresponding definitions. Carroll (1963, 1989) considered the variables of *aptitude, opportunity to learn,* and *perseverance* as being responsive directly to time, in that they focus on how much time a student requires in learning a given task. The variable of aptitude, for instance, considers the amount of time required to learn a task, given typical instruction. Carroll (1963) postulates that high-achieving students, or ones with high aptitude, require less time to learn, whereas students with low aptitude require more instructional time to learn a given task.

TABLE 1. Carroll's (1963) Model of School Learning

	Variables of learning	
Category	Variable	Definition
Within-the-learner	Aptitude	Amount of time a student needs to accomplish a learning task
	Ability to understand instruction	Combination of a student's general intelligence and verbal ability
Outside-the-learner	Opportunity to learn	Time allowed for learning
	Quality of instruction	Organization and presentation of a learning task
	Perseverance	Time the learner is willing to spend actively engaged in a learning task

The second variable, opportunity to learn, is the amount of time prioritized for learning by schools and individual classrooms. Carroll (1989) laments that the amount of time allocated for instruction often falls short of meeting the needs of students. The third variable, perseverance, is the amount of time the individual willingly devotes to learning. In essence, perseverance falls in the hands of the learner. For example, if a student is self-motivated, then he/she will spend the amount of time necessary to learn a given task. In contrast, students who lack interest or motivation in learning often direct too little attention to the task at hand (Kame'enui & Simmons, 1990).

The final two variables, *ability to understand instruction* and *quality of instruction*, associate directly with student achievement. Carroll (1963) defines ability as a student's capacity to gain understanding from instruction even when it is less than optimal. For example, despite poor instruction, students with higher abilities or stronger educational experiences are able to learn on their own. In contrast, students with lower abilities or weaker learning experiences struggle to benefit from instruction that is anything less than ideal. The last variable, quality of instruction, also has direct relevance to achievement. According to Carroll, the quality of instruction is the linkage between teacher behavior and student outcomes. To enhance the quality of instruction, teachers must design and deliver instructional tasks in ways that are accessible for the full range of learners. A focus of this dissertation was to measure the quality of classroom instruction.

Within the context of preventing academic difficulties, Kame'enui and Simmons (1990) divided Carroll's (1963) variables into two separate categories: within-the-learner and outside-the-learner (see Table 1). This separation was timely because it shifted the blame for academic failure from student characteristics (e.g., SES, disability) to more amenable factors that educators can better control, such as the quality of instruction (Simmons & Kame'enui, 1996). In other words, rather than changing the student to fit the system, efforts now emphasize altering the variables that teachers can systematically and strategically manage (Gersten, Baker, Pugach, Scanlon, & Chard, 2001). One variable that lies in the hands of teachers is the quality of instruction.

Measurement of Teaching

Over the course of the past three decades, the measurement of teaching processes has been a cornerstone of education research (Good & Grouws, 1979; Shavelson, Webb, & Burstein, 1986). Results from the study of instruction have played an influential role in generating a knowledge base of effective pedagogical practice (Brophy & Good, 1986; Rosenshine, 1997). Recently, researchers have begun to investigate the quality of instruction with a specific focus on the instructional interactions that occur between teachers and their students (Pianta, 2007; Pianta & Hamre, 2009; Smolkowski & Gunn, 2010). Some postulate that the frequency and quality of teacher-student interactions mediate student achievement (Chard et al., 2008; Hiebert & Grouws, 2007).

Researchers have used a variety of quantitative and qualitative methods to investigate the quality of instruction, including large-scale survey instruments (Rowan, Correnti, & Miller, 2002) and instructional logs (Ball & Rowan, 2004; Rowan & Correnti, 2009). Although different measurement methods demonstrate varied strengths, some are more applicable and more powerful in certain research contexts. Snyder et al. (2006) contend the selection of a particular measurement method should greatly depend on the construct of research interest and the method's sensitivity for detecting sources of variance.

One method for collecting information about classroom instruction and student learning is direct observation (Snyder et al., 2006; Vaughn & Briggs, 2003). According to Snyder et al., direct observation is a sensitive method for handling sources of variance such as time, observer effects, and moment-to-moment changes of teacher and student behavior. Researchers can use direct observations of instruction to explore relationships between teaching practice and student outcomes.

Researchers have developed a variety of direct observation instruments to evaluate classroom instruction. For instance, when examining mathematics instruction, researchers have implemented frequency count approaches (Good & Grouws, 1979), rating scale methods (Clements & Sarama, 2008; Mason & Good, 1996; Pianta & Hamre, 2009), and duration recording systems (Gerleman, 1987). Interestingly, few systems have measured how the frequency of teaching practice relates to kindergarten mathematics achievement. The classroom observation instrument studied in this dissertation centers on the quantity of teaching practice. The system's goal is to measure the frequency of instructional interactions that occur between teachers and students during kindergarten math instruction.

Though direct observation is not exempt from measurement error (Suen & Ary, 1989), many refer to it as the "gold standard" approach for capturing what transpires in actual teaching practice (Ball & Rowan, 2004; Hoge, 1985; Medley & Mitzel, 1963; Palardy & Rumberger, 2008; Rosenshine & Furst, 1973; Snyder et al., 2006). Consider four examples of this last statement. First, most would agree that to truly explain behavior one must pay attention to the important things that people do (Baumeister, Vohs, & Funder, 2007). For this reason, direct observation is an effective method for understanding and explaining the instructional behaviors of teachers and students. Second, the method of direct observation can be sensitive enough to detect change and differences in classroom behavior. For example, Smolkowski & Gunn (2010) found classrooms significantly varied in the amount of practice opportunities students received during kindergarten reading instruction. A well-designed direct observation system should capture differences between treatment conditions (Snyder et al., 2006).

A third example is that direct observation is less susceptible to systematic bias than self-report methods. Self-report methods are more at risk to systematic bias because they rely on individuals who are often aware of intervention status. Threats to self-reports can affect participant motivation (or lack thereof) and participants' perceptions of experimental situation (Shadish, Cook, & Campbell, 2002). Snyder et al. (2006) suggest direct observation is less at risk for systematic bias when researchers (a) operationally define behaviors, (b) train observers to criterion reliability, and (c) use ongoing calculation of rater agreement to minimize observer drift.

Lastly, direct observation collects more trustworthy data than other approaches, such as self-report of past behaviors, teacher interviews, questionnaires, and survey instruments (Hoge, 1985). Baumeister et al. (2007) questioned the dependability of self-report information by stating, "people have not always done what they say they have done, will not always do what they say they will do, and often do not even know the real causes of the things they do" (p. 397). While sometimes more expensive, more intrusive, and less feasible than these other approaches, direct observations can potentially provide a more accurate representation of classroom behavior (Medley & Mitzel, 1963; Snyder et al., 2006). Taken together, direct observation of classroom behavior may be the most appropriate method for gathering valid information about the quality of mathematics instruction and the facilitation of student learning.

Purpose of the Study

In recent years, direct observation has drawn increased attention as a proposed method for investigating the quality of classroom instruction (August, Branum-Martin, Cardenas-Hagan, & Francis, 2009; Gersten, Baker, Haager, & Graves, 2005; Pianta & Hamre, 2009; Stuhlman & Pianta, 2009). In early literacy, for example, researchers have used a variety of classroom observation tools to measure how instructional practices relate to reading outcomes (Baker, Gersten, Haager, & Dingle, 2006; Vaughn & Briggs, 2003). However, less documented in the literature is the application of direct observation instruments during beginning mathematics instruction (Sutherland & Wehby, 2001). A review of the literature reveals that few studies have used direct observation to test the relationship between observed classroom-level behaviors and student math achievement. This is surprising given that a successful start in kindergarten can affect subsequent learning in mathematics (Morgan, Farkas, & Wu, 2009). For these reasons, there is need to identify and measure the effective teaching practices that facilitate student learning during this critical period of children's early mathematical development.

The purpose of this dissertation is to validate a direct observation system. By design, the observation measure systematically captures the instructional interactions that occur between a teacher and their students during kindergarten mathematics instruction. To support the use of the instrument in measuring the quality of instruction, I explore three types of validity evidence. In addition, concerned about the consistency of the observation data collected across a number of independent observers, I examine whether observers could reliably use the instrument in classrooms. Toward this end, I address four research questions.

Research Questions

To address the first research question, this study will conduct a content review of the direct observation instrument. This analysis will allow an examination of the content aspect of construct validity. The second question will use an index of interobserver agreement. This approach will allow for an examination of interobserver agreement and consistency of the data documented by the observation instrument.

1. Does the Coding of Academic Teacher-Student Interactions (CATS) observation instrument include evidence of content relevance and representation for its use in capturing instructional interactions during kindergarten mathematics instruction?

Hypothesis 1: Given its empirical alignment with the scientific research on effective math interventions for promoting student mathematics proficiency, features of the CATS observation instrument will demonstrate evidence of content validity. The study will gather this type of evidence through a 12-item online survey completed by external reviewers.

2. Can observers meet a minimally acceptable level of interobserver agreement when using the CATS observation instrument in both experimental classrooms (i.e., where teachers have been trained to teach the Early Learning in Mathematics curriculum; ELM) and comparison classrooms (i.e., where teachers have not been trained)?

Hypothesis 2: Based on comprehensive observer training and the simple structure of the observation instrument, observers will reach a minimally acceptable level of agreement across both treatment and comparison classrooms, and across observation rounds.

To address the final two research questions, this study will employ a hierarchical design and develop multilevel statistical models (two-level) that nest (a) observations within classrooms and (b) students within classrooms. The first model will allow an investigation of the instructional differences of teachers between classroom conditions. The second will allow an examination of the relationship between average rate of observed behaviors and student math achievement.

3. Can the CATS observation instrument detect differences in mathematics instruction provided by treatment and comparison classrooms?

Hypothesis 3: Because the treatment curriculum (ELM) emphasizes explicit instruction and frequent student practice opportunities, I expect the observation instrument will detect differences in the mean rate of instructional behaviors between treatment conditions.

4. Is there a relationship between the average rate of observed classroom-level behaviors, as measured by the CATS observation instrument, and student mathematics achievement? And to what extent do the observed behaviors moderate the relationship between risk status and student mathematics achievement?

Hypothesis 4: Based on a growing body of empirical evidence on the effective principles of instruction, the analysis will detect a statistically significant relationship between the observed behaviors and student mathematics achievement. Moreover, I hypothesize that the observed behaviors will moderate the relationship between risk status (as classified by utilizing a 20th-percentile cutoff on two curriculum-based pretest measures) and student math achievement.

CHAPTER II

A REVIEW OF THE LITERATURE

The literature review for this study centers on two research-based principles of instruction. These principles are practice opportunities and explicit instruction. An emerging body of scientific research has shown that both instructional principles are an effective approach for improving math outcomes for students with and without mathematics learning disabilities. It is also important to note that each principle maps directly onto the classroom observation instrument studied in this dissertation. I provide a thorough description of the observation instrument in Chapter III. To shape this literature review, I draw upon the findings of three relevant sources: (a) studies from the process-product literature, (b) recent experimental and quasi-experimental studies that test methods for explicitly teaching mathematics to students with or at risk for difficulties, and (c) meta-analyses for teaching mathematics to students with learning disabilities.

Practice Opportunities

Children enter kindergarten with varying levels of mathematical understanding (Anuola, Leskinen, Lerkkanen, & Nurmi, 2004). Some children, for instance, can rote count from 1 to 20, solve simple plus-one addition problems, and extend basic ABAB patterns. Other children have difficulty identifying basic numerals. Why do these differences in math readiness exist? One plausible answer is that children receive different levels of support in their home environments for building early mathematical knowledge.

Prior research has found the difference in informal learning experiences to be socioeconomic status (SES)-related (Bodovski & Farkas, 2007; Denton & West, 2002; Guarino et al., 2006; Klein, Starkey, Clements, Sarama, & Iyer, 2008; Rathbun & West, 2004; Sarama, Clements, Starkey, Klein, & Wakely, 2008; Starkey & Klein, 2000; West et al., 2001). Children from higher SES households receive higher levels of support for early math development prior to entering kindergarten than their economically disadvantaged peers. Thus, by the time economically disadvantaged children enter school, they have received fewer opportunities to engage in number-related activities (Klein et al., 2008). Because of this variation, it is important to provide frequent and rich practice opportunities in early mathematics to kindergarten children with lower math knowledge.

Well-designed practice opportunities benefit both students and teachers. For students, practice helps build conceptual and procedural knowledge (Miller & Hudson, 2007; NRC, 2001). Also, regular practice allows students to maintain newly acquired skills (Tournaki, 2003) and shift to memory-based retrieval of number combinations (Fuchs et al., 2010). For teachers, practice opportunities can help them gain critical information about their students. For example, teachers can better estimate individual and group understanding when they provide students with frequent opportunities to learn (Carnine, Silbert, Kame'enui, & Tarver, 2004; Harniss, Carnine, Silbert, & Dixon, 2007).

In mathematics instruction, practice opportunities take on many shapes. Classrooms, for instance, use computerized practice activities, games, written exercises, and discussion to build skill fluency. One of the most common forms of practice is textbook and worksheet exercises. These learning opportunities typically entail students working on newly acquired skills, absent of teacher guidance. When purposefully distributed, worksheet exercises offer students with opportunities to practice new and previously learned skills (Carnine, 1997). For example, well-designed instructional programs provide concentrated practice following the introduction of a new skill and systematic review for maintaining new material (Carnine et al., 2004; Chard & Jungjohann, 2006).

To extend student understanding, practice opportunities must also include activities that extend beyond the typical drill and practice format of written exercises. For example, activities must allow students to engage in systematic opportunities of mathematics-related dialogue and application of key mathematics concepts and principles. Teachers can promote these types of practice activities by engaging their students in effective instructional interactions.

Instructional interactions between teachers and students are an integral part of mathematical learning (Cohen, Raudenbush, & Ball, 2003; Hiebert & Grouws, 2007; Shuell, 1996). These interactions can be bidirectional. For instance, an instructional interaction might consist of a group of students verbalizing their solution for solving a word problem. Another instructional interaction might consist of a teacher model demonstrating the physical attributes of a rhombus and then asking an individual student to count the sides of the shape (see Figure 1). Using the same example, the teacher might ask the whole-class to hold up as many fingers as sides of the shape.

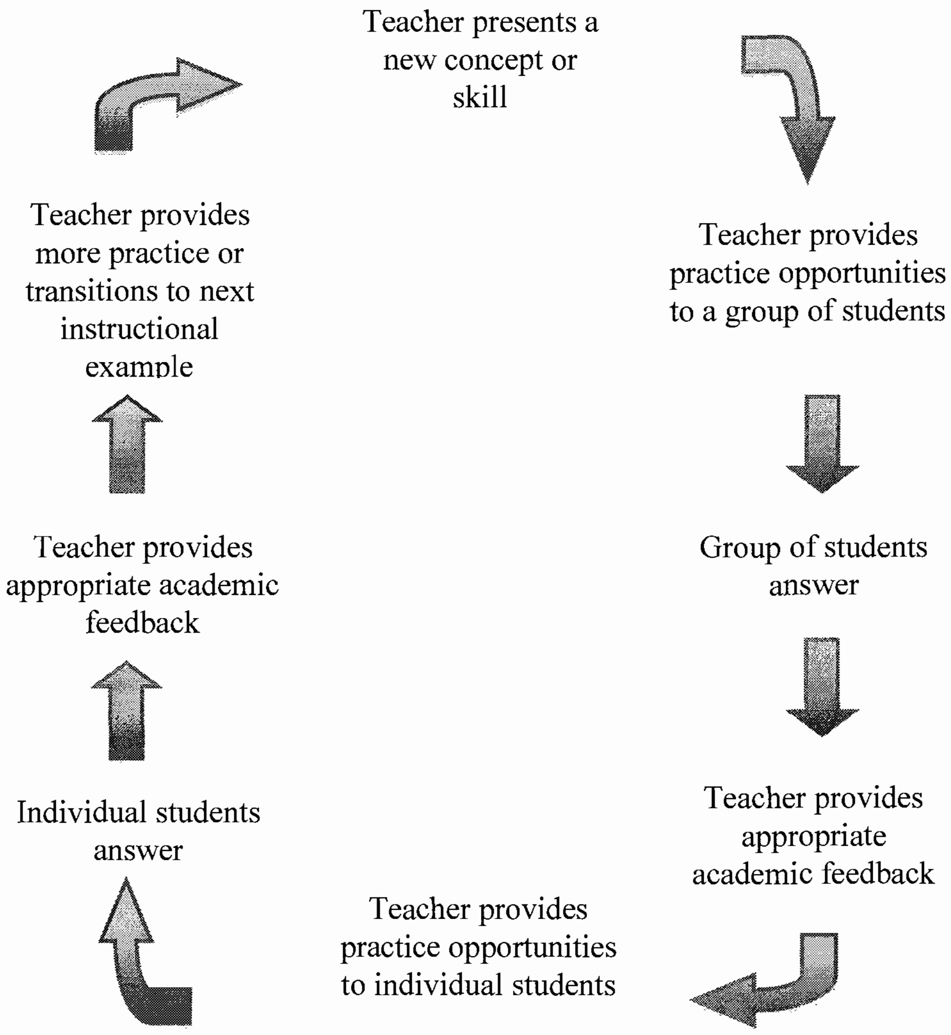

Teacher presents a new concept or skill
Teacher provides practice opportunities to a group of students
Group of students answer
Teacher provides appropriate academic feedback
Teacher provides practice opportunities to individual students
Individual students answer
Teacher provides appropriate academic feedback
Teacher provides more practice or transitions to next instructional example

FIGURE 1. An example of an instructional interaction between a teacher and their students.

Figure 1 presents an example of an instructional interaction in which a teacher (a) models a new mathematics concept, (b) provides practice opportunities at the group and individual levels, and (c) extends student learning with timely, academic feedback.

One aspect of an instructional interaction that maximizes practice opportunities for a group of students is unison oral responding. This type of responding involves a simultaneous response from two or more students (Carnine et al., 2004). According to Carnine et al., unison oral responding is beneficial for several reasons. First, a unison response can facilitate student learning for a group of students. Second, a unison response can keep a group of students actively engaged. Finally, a unison response can provide the teacher with frequent checks to gauge student understanding. In the context of teaching beginning reading, Carnine et al. recommend providing between 10 and 15 response opportunities per minute. In mathematics, teachers are likely to meet this response rate during numeral identification activities and fluency practice with basic addition and subtraction number combinations.

Over the last 30 years, the use of practice opportunities has become a focal point of several research studies. One of the better known works is the Missouri Math study conducted by Good and Grouws (1979). The study examined the teaching behaviors of 40 classroom teachers, randomly assigned to either treatment or control conditions. Teachers delivered instruction for both treatment and control conditions in whole-class format. While control conditions consisted of business-as-usual instruction, treatment instruction incorporated frequent practice opportunities through daily review, seatwork, homework, special reviews and monthly reviews. Good and Grouws based these practice opportunities on the active teaching model (Brophy & Good, 1986). In efforts to maximize the effect of instruction, researchers trained the treatment teachers to keep instruction at a brisk pace.

To test their hypotheses, Good and Grouws (1979) observed each teacher six times. Observation findings indicated that treatment teachers exhibited more behaviors related to the active teaching model than control teachers. For instance, observers noted that treatment teachers spent more time practicing and reviewing previously learned concepts and skills. While Good and Grouws expected these findings based on the training that treatment teachers received, it is important to note that the experimental model had a significant effect on the achievement of students in the treatment condition. Interestingly, despite initial differences of achievement scores favoring the control group, results indicated statistical differences in favor of the treatment group. Using a mean comparison of raw scores, students in the treatment group gained over three points more than their control group peers on a standardized math assessment.

Brophy (1999), in a review of the process-product literature, synthesized a series of 12 instructional principles associated with promoting student achievement. Among the principles was the use of practice and application activities. Under the notion that practice leads to mastery learning, Brophy emphasized its role in teaching complex concepts and skills. Brophy suggested that teachers help students learn through three different means of instruction. First, teachers explain concepts, and demonstrate skills and strategies. Second, teachers ask questions and initiate classroom discourse. Third, teachers provide students practice opportunities to build fluency and maintain knowledge. To help students maintain the information they learn, Brophy suggests that teachers need to provide well-structured practice opportunities. According to Brophy, it is best to avoid trial and error practice. To successfully build students' skills, practice must be systematic and efficient. Moreover, it must extend beyond the redundancy of fill-in-the-blank worksheets and include more application activities.

In a recent random control trial, Fuchs et al. (2008) examined the influence of two practice-oriented interventions on the outcomes of 133 third-grade students with math disabilities. The first intervention, Math Flash, was a computerized practice activity purported to build students' fluency in solving basic number combinations (e.g., 6 + 3). Each Math Flash lesson lasted between 20 and 25 minutes. The second intervention, Pirate Math, targeted students' problem-solving skills. Similar to Math Flash, Pirate Math contained lessons lasting 25-30 minutes. Random assignment placed students in one of three conditions: the Math Flash intervention, the Pirate Math intervention, or a comparison group. Findings of the study were quite convincing, favoring both of the treatment conditions. For example, when compared to the control group, the Math Flash effect size was large, 0.85. Though smaller than the Math Flash effect, the effect size for Pirate Math was moderate to large, 0.72. Fuchs et al. (2008) postulated the effect size differences between the interventions were attributable to fewer practice opportunities offered in the Pirate Math condition.

Explicit Instruction

There is converging evidence that explicit instruction is the most effective method for teaching students with or at risk for math disabilities (Baker, Gersten, & Lee, 2002; Darch, Carnine, & Gersten, 1984; Gersten, 1985; Gersten et al., 2009; Haas, 2005; Jayanthi, Gersten, & Baker, 2008; Kroesbergen & Van Luit, 2003; NMAP, 2008; White, 1988). Often referred to as direct instruction, systematic and explicit instruction draws its early roots from the work of Siegfried Engelmann and colleagues (Becker, Engelmann, Carnine, & Rhine, 1981) and Barack Rosenshine (1979, 1983). Carnine et al. (2004) state, "Direct instruction involves teaching . . . essential skills in the most effective and efficient manner possible" (p. 5). Within the explicit teaching model are several critical features, including teacher demonstration, guided practice, and academic feedback (Rosenshine & Stevens, 1984). A hallmark of direct or explicit instruction is its insistence of learning for mastery (Gersten, 1985). According to the direct instruction model, it is imperative that students master each and every step in the learning process before proceeding in the instructional material.

For teachers, explicit instruction entails directly providing students with clear explanations and timely academic feedback. For students, explicit instruction provides ample opportunities to respond and practice. These opportunities include answering and asking questions, verbalizing problem-solving solutions, and completing guided practice tasks. In most cases, instruction concludes with students independently completing a cumulative review of new and previously learned material. During the review tasks, teachers frequently monitor for student understanding.

One of the many notable strengths of the explicit instructional approach is that it centers on high rates of student success (Gersten et al., 2001). Teachers can ensure higher levels of success by modeling and communicating clear strategies during initial instruction (Carnine, 1997; Harniss et al., 2007; Hudson & Miller, 2006). For example, in the context of teaching a multistep math procedure, teachers can facilitate student learning by overtly presenting and describing the steps for solving the problem.

The research base in support of explicit instruction for teaching students with learning difficulties is sound. Recent meta-analyses have provided much of this empirical backing. For instance, in an early meta-analysis, White (1988) reviewed 25 studies employing a direct instruction approach. Findings of the analysis revealed that students with learning disabilities demonstrated stronger outcomes in the direct instruction interventions. Using a mean effect size calculation, White found large effects (0.82) in favor of direct instruction. Swanson and Hoskyn (1998) also completed a thorough review of the direct instruction literature, analyzing the effects of 180 research studies. Similar to White's (1988) findings, Swanson and Hoskyn (1998) found direct instruction one of the most effective teaching approaches. Calculated effect sizes revealed a moderate to large impact (0.68) on student achievement.

In a more recent meta-analysis, Baker et al. (2002) reviewed the literature on teaching mathematics to low-achieving students. Baker et al. used the term "low-achieving" instead of "learning disabled" to include more students at risk for math failure. Using specific inclusion criteria, the meta-analysis yielded a total of 15 experimental and quasi-experimental intervention studies. Baker et al. codified the 15 studies according to five instructional categories. One category focused on the use of explicit instruction and contextualized teacher-facilitated instruction. Seven studies met the instructional category criteria. Calculating effect sizes through standardized mean differences, findings revealed moderate to large effects for the explicit instruction (Average = 0.65, Weighted = 0.58). In contrast, the teacher-facilitated approach yielded negative to no effects on student achievement. These findings lend further support for the use of explicit instruction when teaching mathematics to struggling learners.

Kroesbergen and Van Luit (2003) conducted a meta-analysis involving 58 studies of mathematics interventions for students with learning disabilities. In the study, Kroesbergen and Van Luit categorized interventions into one of three intervention domains: preparatory mathematics, basic skills, and problem-solving strategies. One goal of the meta-analysis was to determine if treatment components, such as direct instruction, self-instruction, computer-assisted instruction, and peer tutoring, were effective for improving student math outcomes. At the basic skills domain, Kroesbergen and Van Luit found direct instruction the most effective, revealing a large weighted effect size of 1.13.

The National Mathematics Advisory Panel (NMAP, 2008) made a recent attempt to conduct a meta-analysis of instructional approaches for teaching mathematics. Because the literature base is “not uniformly deep” (p. 6-1), the Panel was unable to take on a meta-analytic approach for comparing student-centered and teacher-directed instruction. The Panel, however, did make several recommendations for teaching low-achieving students and students with learning disabilities, based on 26 high-quality experimental studies. Findings from the review revealed explicit instruction as the most appropriate approach for improving struggling learners' performances in computation and word problems.

Bryant et al. (2008) used an explicit instructional approach in a recent regression-discontinuity study. Using a 25th percentile cut-score on a standardized math assessment (Texas Early Mathematics Inventories: Progress Monitoring, TEMI-PM), Bryant et al. identified 161 Grade 1 students eligible for the Tier 2 intervention. Students who scored at or above the cut-score received typical math instruction. Within the Tier 2 treatment, students received explicit instruction in key concepts and skills. Topics included in the study were counting, number sense, place value, and basic operations. Treatment included 20-minute tutoring sessions offered four days per week, using instructional features such as pacing, opportunities to respond, error correction, and strategy instruction. Bryant et al. (2008) found a significant main effect for the intervention at the conclusion of the treatment.

Most recently, Gersten et al. (2008), using meta-analysis, examined the effects of interventions for teaching math to students with learning disabilities. Unlike Baker et al. (2002), Gersten et al. (2008) chose to include only those studies that involved students with learning disabilities. The search of the literature yielded 42 interventions. Gersten and colleagues categorized the studies into four categories, one of which was instructional approaches. This category included the approach of explicit instruction.

Because researchers interpret the concept of explicit instruction in various ways, Gersten et al. applied three specific inclusion components. First, interventions had to include a step-by-step plan for solving problems. Second, the step-by-step plans had to map onto the problem type targeted during instruction. Third, students had to apply the same step-by-step procedure that the teacher previously demonstrated. At this category level, the meta-analysis examined 11 studies. The mean effect size of explicit instruction was substantively large, 1.22. The findings of Gersten et al. (2008) corroborate the results of earlier meta-analyses (Baker et al., 2002; Kroesbergen & Van Luit, 2003; Swanson & Hoskyn, 1998; White, 1988). However, despite the robust results favoring explicit instruction, Gersten et al. (2008) were quick to remind the reader that it is important to consider a mix of instructional methods when teaching different concepts and skills.

Summary and Linkages to Present Study

Carroll's (1963, 1989) model postulates five variables associated with school learning. Of the five variables, *opportunity to learn* and *quality of instruction* were most relevant to this dissertation. Children require opportunities to acquire new knowledge. Perhaps above all, children need learning opportunities that are of high quality and effective in structure. Previous research indicates that struggling learners best acquire mathematical concepts and skills when taught with research-based principles of instruction, such as structured practice opportunities and explicit instruction. When taught in combination, these principles can help promote effective instructional interactions between a teacher and their students. For these reasons, there is need to document the quality and quantity of instructional interactions that take place in kindergarten math classrooms. A direct observation system may be the most applicable and most powerful method to address this need.

Thus, the objective of this dissertation was to validate a direct observation instrument designed to capture the frequency of student practice opportunities, teacher demonstrations, and teacher-provided academic feedback. Some hypothesize that when instructional interactions contain these principles of instruction they facilitate learning for all students in general and students with diverse learning needs in particular.

CHAPTER III

METHODOLOGY

The purpose of this dissertation was to validate the Coding of Academic Teacher-Student (CATS), a direct observation instrument that systematically measures the instructional interactions that occur between teachers and students during kindergarten mathematics instruction. In conducting this study, I addressed a set of research questions related to three types of validity evidences. First, the study tested for evidence of content relevance and representativeness (e.g., content validity; Messick, 1995). Second, it assessed whether the instrument was sensitive to detect differences in instruction between two types of classroom conditions (e.g., discriminant validity). Third, it tested the relationship between instructional interactions, captured by the instrument, and end-of-year math outcomes for kindergarten students (e.g., criterion-predictive validity). In addition, the study examined the consistency of the observation data collected across a number of independent observers (e.g., interobserver agreement). By addressing these particular questions, this study sought to establish a valid and reliable observation instrument, and explore the instructional practices hypothesized to directly influence student learning.

Design

This study employed a hierarchical design and developed several multilevel statistical models. To address the study's four research questions, data were harvested from an ongoing randomized control efficacy trial called the Early Learning in Mathematics: Efficacy Trials in Kindergarten Classrooms (ELM-ETKC; Baker, Chard, Clarke, Smolkowski, & Fien, 2008). These data contained both student and classroom-level information. Because the observation instrument focuses on the instructional behaviors both directed and supported by teachers, the primary unit of analysis for this dissertation was classrooms.

The ELM-ETKC Project

The Early Learning in Mathematics: Efficacy Trials in Kindergarten Classrooms project (Baker et al., 2008), hereafter referred to as ELM-ETKC, is a randomized controlled trial that is investigating the efficacy of the Early Learning in Mathematics (ELM) kindergarten curriculum. In addition to testing the immediate and long-term impact of the ELM curriculum on student mathematics achievement, the ELM-ETKC project is also systematically investigating the mediating and moderating variables hypothesized to influence student learning across conditions. The project recently completed its first year of investigation.

Year 1 of ELM-ETKC involved 65 classrooms from three school districts across the state of Oregon. Two of the school districts are located in the suburban area of

Portland and one in the southern part of the state. Across the three school districts, student ethnicity ranged as follows: Caucasian (54% to 74%), African American (1% to 2%), Latino (18% to 30%), Asian/Pacific Islander (2% to 7%), Native American/Alaskan Native (1% to 2%), and other ethnicities (0% to 5%). The percentage of students receiving special education services across the districts ranged from 10% to 14%. Student enrollment in free or reduced lunch programs ranged from 32% to 47%.

Of the 65 participating classrooms, the project randomly assigned 35 within schools to the treatment condition (ELM) and 30 to the comparison condition. Instruction in the treatment condition consisted of the ELM curriculum using a whole-class instructional format. The project considered the comparison condition as "business as usual" practice. Comparison classrooms employed instructional materials approved by their respective district.

Participants

A total of 66 teachers in 65 classrooms participated in Year 1 of the ELM-ETKC project. All 66 teachers remained in the study throughout the intervention year (0% attrition). Most teachers were females (97%). The two male teachers involved in the study taught in treatment classrooms. Participating teachers had an average of 10.52 years of teaching experience, and a mean of 6.35 years teaching at the kindergarten level. Across conditions, 56% of the teachers held a graduate degree, and 88% identified as Caucasian, 8% Hispanic, and 5% other ethnicities. Approximately 13 (43%) of the comparison teachers and 17 (49%) of the treatment teachers completed college-level coursework in Algebra.

Of the 65 participating classrooms, the majority of classrooms involved one participating teacher. In one classroom, however, two teachers worked a full-time equivalent of 0.5 or half-time. Forty-eight classrooms provided a full-day kindergarten program and 17 offered a half-day program. Of the 17 half-day classrooms, 10 were in the treatment condition. One classroom offered a full-day program four days per week, while the remaining 64 classrooms offered instruction five days per week. Average class size for the treatment and comparison conditions was 19.5 and 19.4, respectively. Each classroom contained approximately 12 boys. Participating classrooms received assistance from instructional aides an average of 0.60 hours per week.

Nested within the 65 kindergarten classrooms were approximately 1,495 students. Of the participating students, approximately 92 dropped out of the study (6.1% attrition) primarily because of family mobility. Approximately 77 students moved into the participating classrooms during the 2008-2009 school year.

On an end-of-year administered survey, teachers reported an average of 11 students per classroom as at risk for failure in mathematics. Teachers based this identification on student performance from the 2009-2009 school year. On the same survey, teachers reported an average of 9.04 English language learners in each classroom. The percentage of students receiving special education services across the classrooms ranged from 1.26% to 1.39%.

Dissertation Sample

This dissertation utilized an existing data set provided by the ELM-ETKC project (Baker et al., 2008). The data set included student and classroom-level information collected in the 65 kindergarten classrooms participating in ELM-ETKC. Data accessed at the student level included pretest scores from two curriculum-based measures (Oral Counting, Number Identification; Clarke & Shinn, 2004), and pretest and posttest scores from a standardized mathematics outcome measure (Test of Early Mathematics Ability-Third Edition [TEMA-3]; Ginsburg & Baroody, 2007). Collection of student-level data took place in fall (October-November 2008) and spring (April-May 2009) of the kindergarten school year.

Across the 2008-2009 school year, the ELM-ETKC project involved 1,495 students. From this sample, ELM-ETKC assessed approximately 1,200 (80%) in the fall (pretest) and 1,246 (83%) in the spring (posttest). The analytic sample in this study included 929 kindergarten students. This sample included only those students who participated in the fall and spring administrations of the outcome measure (TEMA; Ginsburg & Baroody, 2007) and the pretest administrations for both curriculum-based measures (Oral Counting and Number Identification; Clarke & Shinn, 2004).

At the classroom level, I accessed observation data, and demographic information related to teachers and classrooms. The ELM-ETKC project collected demographic information in the fall of the intervention year. The observation data included quantitative information relevant to explicit instructional practices such as the frequencies of teacher models, academic feedback, and student practice opportunities.

The ELM-ETKC project planned three direct observations per classroom across the fall, winter, and spring of the 2008-2009 school year. Approximately 6 weeks separated each observation round. Each round planned for one observation per classroom for a total of 65 observations per round. In all, ELM-ETKC conducted a total of 191 observations. The first round (fall) involved 62 observations, while the second (winter) and third (spring) rounds involved 64 and 65 observations, respectively. It is important to note that classrooms were observed no more than one time per observation round. Of the 65 classrooms, 61 were observed at each observation round (i.e., fall, winter, and spring). Four classrooms were observed on just two different occasions across the observation rounds. Missing observations (i.e., <3% of scheduled observations) were primarily due to scheduling conflicts or teacher absences.

Trained observers conducted all observations. Classroom observations took place during the core mathematics instruction time period. Due to random assignment within the ELM-ETKC project, core math instruction for treatment classrooms consisted of whole-group instruction in the Early Learning in Mathematics curriculum. Instructional formats (i.e., small-group and whole-class) and teaching materials varied across comparison classrooms.

Of the 191 observations, the ELM-ETKC project conducted approximately 24% (n = 46) as paired observations. A paired observation consisted of two observers independently measuring kindergarten mathematics instruction. For research studies involving direct observation, these types of pairings are critical for monitoring whether there is consistency of data collection across a number of independent observers throughout a given time period (Kennedy, 2005; Shoukri, Asyali, & Walter, 2003; Suen & Ary, 1989). Some postulate that demonstration of interobserver agreement reduces possible sources of bias, measurement error, and variance attributable to characteristics of independent observers (Baker et al., 2006; Brennan & Johnson, 1995; Kennedy, 2005; Messick, 1995; Parkes, 2007; Raudenbush & Sadoff, 2008; Snyder et al., 2006; Suen & Ary, 1989). In the case of this dissertation, I anticipated that the interobserver agreement data would provide initial support for using the instrument to measure instructional interactions in kindergarten math classrooms.

Coding of Academic Teacher-Student Interactions (CATS) Observation Instrument

This dissertation centered on the Coding of Academic Teacher-Student interactions (CATS) observation instrument. The CATS tool systematically measures the instructional interactions that occur between teachers and students during kindergarten math instruction, such as student practice opportunities and teacher demonstrations. It is important to note that members of the ELM-ETKC project, myself included, developed the CATS tool specifically for use in the efficacy study. However, given its conceptual alignment with the effective principles of instruction, the ELM-ETKC team suspects the instrument's use will generalize to other educational contexts, such as measuring the instructional interactions of kindergarten reading instruction.

The CATS observation tool was designed to assess effective instruction for early mathematics learning (see, e.g., Gersten et al., 2009; Jayanthi et al., 2009; NMAP, 2008). Certain features of the CATS were adapted from the Student-Teacher Interactions Context Observation instrument (STICO; Smolkowksi & Gunn, 2010). Smolkowski and Gunn recently used the STICO instrument to measure instructional interactions during kindergarten reading instruction. Results of the STICO study indicate that the rate of student practice opportunities was a significant predictor in early reading outcomes.

General Features of the CATS

The CATS uses an event or frequency recording system to collect information about classroom instruction. As such, the instrument requires observers to document each time an instructional behavior occurs. Three sections comprise the CATS: (a) cover page for general information about the observation occasion, (b) Context Code section, and (c) Instructional Interaction Code section. Appendix A presents a complete copy of CATS.

The cover page for CATS requires observers to record general information about the observation, including identification numbers for the school and observed teacher. Also included on the cover sheet are the observation start and stop times, total number of students in the class, math program and lesson number taught during the observation, group size, date of observation, and the observer's initials.

The context code section entails four components: (a) instructional start and finish times for the math activity, (b) type of math content targeted in the activity, and (c) type of instructional format (small-group or whole-class). The mathematical content areas include (a) number and operations, (b) geometry, and (c) measurement. The ELM-ETKC project chose to capture these particular areas because of their prominent role in the kindergarten curriculum (Clements, 2004; National Council of Teachers of Mathematics [NCTM], 2006; NMAP, 2008). There is also strong consensus among researchers, math educators, and expert panels that early achievement in mathematics requires students to develop proficiencies in an array of concepts and skills associated with these three content areas (Clements, 2004; NRC, 2001; Van de Walle, 2001; Wu, 2001).

During an observation, observers code one content area per instructional activity. When a teacher introduces a different math activity, observers would then code a new content area. For example, if the teacher begins the lesson with a numeral identification activity and then transitions to recognizing shapes, the observer would code the first activity as number and operations, and the second as geometry. The observer would also note the stopping time of the previous activity and the starting time of the new activity. Then the observer would begin with a new coding sequence for the current activity.

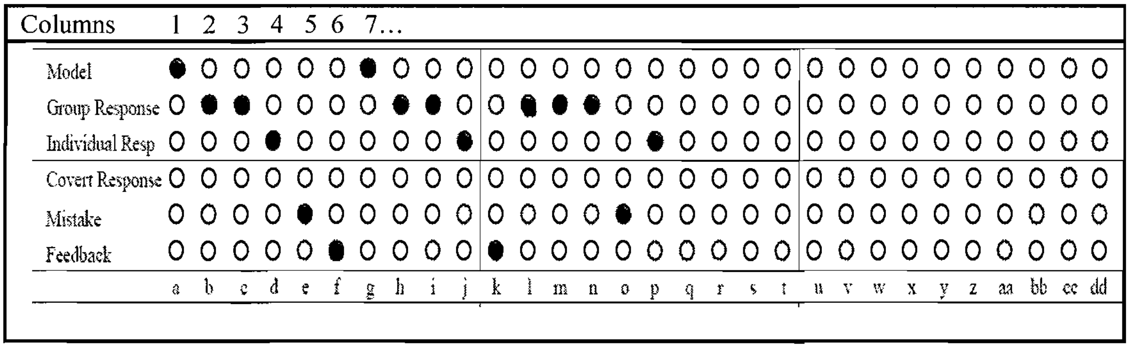
If the teacher transitions to a non-math-related activity, such as a brief reading activity or an instance of classroom management, the observer would code the activity as a nonmath content area, or “other,” and complete the same context code components as for a math activity. This includes noting the start and stop times of the adjacent activities, and the type of instructional format. Because the focus of the ELM-ETKC project is mathematics instruction, observers do not record the instructional interactions that occur during nonmath activities.

Instructional Features of the CATS

The instructional interaction section focuses on six behaviors: (a) teacher models, (b) group responses, (c) individual responses, (d) covert responses, (e) student mistakes, and (f) teacher-provided academic feedback. Observers code behavior occurrences in a continual, serial fashion. Using a frequency count method allowed this study to measure the relationship between the quantity of teaching, such as the rate of individual practice opportunities, and student learning. It is important to note that analyses for this dissertation concentrate on individual and group responses, and teacher models.

Figure 2 presents an example of a coding sequence, with the first column illustrating an initiating teacher model, followed by two group responses in columns 2 and 3. Columns 4 and 5 indicate an individual response that was incorrect. As displayed in column 6, the teacher recognized the student's error and provided some degree of academic feedback. Figure 2 also illustrates that the observer correctly recorded one behavior code per column.

The CATS measures two teaching behaviors, with the first being teacher model. A teacher model is an explicit and overt teaching behavior. Teacher models include explanations, verbalizations of thought processes, and physical demonstrations. To be considered as a codable behavior, teacher models must focus on mathematical content. For example, observers would code a model if a teacher states a math definition or demonstrates a multistep mathematical procedure. In contrast, observers would not code a model if a teacher demonstrates where students should write their names on a math practice worksheet.



Columns 1 2 3 4 5 6 7...

Model	●	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Group Response	○	●	●	○	○	○	○	●	●	○	○	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Individual Resp	○	○	○	●	○	○	○	○	○	●	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Covert Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mistake	○	○	○	○	●	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Feedback	○	○	○	○	○	●	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	bb	cc	dd

FIGURE 2. A coding example of an instructional interaction.

The second teaching behavior is academic feedback. Academic feedback is an overt teaching behavior that specifically relates to (and immediately follows) a group or individual student response. Academic feedback can take the form of either an error correction or a response affirmation. Both forms can include a teacher's verbal reply, physical demonstration, or written response. Observers code academic feedback when a teacher corrects a student mistake or when the teacher affirms a correct student response.

The CATS measures four student behaviors: (a) group responses, (b) individual responses, (c) covert responses, and (d) mistakes. As with the teacher behaviors, observers used similar processes to code the student behaviors. Observers code each time a student behavior occurs. The first behavior is group response, which consists of a verbal response from two or more students. A group response, for example, would consist of 15 students concurrently stating the name of a three-dimensional shape. Accurately coding physical demonstrations and written responses from more than one student can be difficult. Therefore, these types of behaviors are not included under the coding scheme of group response. Instead, the CATS tool requires observers to code these types of group behaviors as covert responses.

In contrast, an individual response, the second student behavior, includes verbal responses or explanations, physical demonstrations, and written answers. Individual responses can be elicited from students specifically identified by the teacher (e.g., "Johnny, what shape?") or questions posed to the group at large (e.g., "Who can count to 5?"). For both group and individual responses, the tool requires a teacher-posed question and/or request to precede a student answer. Thus, this avoids observers coding student callouts or extraneous conversation.

The third student behavior is student mistake. Errors can involve verbal mistakes (e.g., a counting mistake) and physical mistakes (e.g., pointing to an incorrect shape). Moreover, errors can occur both at the group and individual level. For example, an observer would code an error if one student made a mistake during a group counting activity. Observers code incorrect responses only when the error is clearly visible (or audible) to the teacher and the observer.

The fourth and final student behavior is covert response. A covert response is a physical demonstration of math learning. It's important to note that a covert response is an observable behavior. However, unlike the group and individual responses, a covert response is a behavior that is difficult to capture from the observer's vantage point. Covert responses include group written responses, use of counting fingers or math models by two or more students, and partner learning. An example of a covert response during math instruction is peer partners counting by fives to 50. In this case, given the difficulty of tracking both correct and incorrect responses from student pairings, observers would code a covert response. Another example of a covert response is 18 students holding up three fingers to represent the numeral three.

Observer Training and Interobserver Agreement

Prior to each observation round, the ELM-ETKC observation team provided comprehensive training in the CATS and the procedure of direct observation. Training for the first round of observations consisted of 12 hours of preparation. To minimize observer drift, the observation team also provided 4-hour refresher training sessions prior to the second (winter) and third (spring) rounds of observations. Collectively, the observers received over 20 hours of training in (a) direct observation procedures, (b) kindergarten mathematics content, and (c) the CATS observation instrument.

Video Checkouts

At the conclusion of each training session, observers were required to complete two reliability checkouts. The first was a video checkout, which required observers to (a) watch a 5-minute ELM activity, filmed in a pilot study classroom; and (b) code all observable behaviors. The video contained a total of 70 code-able behaviors. Observers were required to meet an interobserver agreement level of .80 for coding all behaviors contained on the video (i.e., overall agreement of combined teacher and student behaviors). Observers were also required to meet an agreement level of .80 for the categorized behaviors, including combined teacher behaviors (i.e., teacher models, academic feedback), and combined student behaviors (i.e., group and individual responses, covert responses, and student mistakes). The observation team selected a criterion of .80 because it met the minimal level of interobserver agreement recognized by traditional and current standards (e.g., Horner et al., 2005).

To calculate observer agreement for all video checkouts, the observation team used a frequency-ratio approach (Kennedy, 2005). The frequency-ratio approach, also known as a smaller/larger index (Suen & Ary, 1989), takes the sum of observed behaviors from two observers and divides the smaller value by the larger value. Resulting agreement values range from 0.00 to 1.00. For the agreement calculations, the observation team considered the answer key from the video as the second observer. Trainees completed the video checkout until they met the 80% agreement cutoff.

Classroom Checkouts

In addition to the video checkout and prior to heading out into the field on their own, observers were required to complete a real-time classroom reliability check. Classroom checkouts consisted of two observers: one primary observer, or "the standard," and one secondary observer, or trainee, completing concurrent data collection. There were three primary observers for the ELM-ETKC project, including one of the project's principal investigators, the observation coordinator, and the author of this dissertation. Trainees, when compared to a primary observer, were required to meet the same agreement levels as described in the video checkouts (i.e., ≥ .80 for overall agreement, ≥ .80 for combined teacher behaviors, and ≥ .80 for combined student behaviors). To calculate interobserver agreement for the classroom checkouts, the observation team also applied the frequency-ratio approach or smaller/larger index (Suen & Ary, 1989). If trainees fell below the reliability cutoff, they were required to complete a second observation pairing. Once trainees met the classroom checkout, they were clear to conduct observations on their own.

A total of ten observers, nine trainees and one observation coordinator, received the observation training during Year 1 of the study. All 10 observers passed both the video and classroom reliability checkouts (interobserver agreement range = .80 to 1.00). Of this group, five collected observation data at each observation time period. Both the author of this dissertation and the principal investigator completed classroom observations in the three observation rounds. Taken together, the observation team included a total of 12 independent observers across Year 1. The team included three former elementary school teachers, a fourth-year doctoral student studying school psychology, a research associate at the University of Oregon, and seven data collectors from a nonprofit research institute.

Kindergarten Mathematics Instruction

ELM Classrooms

Instruction in the 35 treatment classrooms entailed the Early Learning in Mathematics (ELM) curriculum. Treatment teachers delivered the ELM curriculum using a whole-class instructional format. The teachers in the 35 ELM classrooms taught mathematics in English. The ELM curriculum employs an explicit instructional approach for teaching (a) mathematical representations, (b) math-related vocabulary and discourse, and (c) procedural fluency and mastery of key concepts and skills. The program is comprised of 120 forty-five-minute lessons, with an additional 15-minute calendar activity. Every fifth lesson of the program integrates a problem-solving activity. Most lessons entail four to five activities across the mathematical strands of number and operations, geometry, and measurement. Lessons contain explicit teaching examples, and practice opportunities to review new and previously taught material. In order to complete the entire curriculum in one school year, treatment teachers are supposed to teach one lesson per day, five days per week.

Comparison Classrooms

For the 30 comparison classrooms, teachers used a variety of commercially available and teacher-developed mathematics materials. For example, some classrooms employed the kindergarten edition of the Everyday Mathematics program, while others used the Houghton Mifflin Math program. Of the 30 comparison classrooms, one used Spanish for teaching mathematics, while the remaining classrooms used English. Comparison teachers also taught mathematics using a host of instructional formats, including whole-class instruction and center-based learning. Center-based learning is the formation of small groups, usually of varying student ability, working on different concepts or skills. Center group sizes typically range between three and six students. During the instructional time period, groups can (a) transition between centers, working on different activities at each site; (b) remain at one center for the entire block of math instruction; or (c) begin with whole-class instruction and transition to center-based learning.

Measures

In the following section, I provide a brief description of the predictor variables and measures used in the dissertation. In all, the study employed a total of six measures. With regard to the predictor variables, I examined characteristics of both students and classrooms.

Student-Level Predictors

To account for differences of student performance at time of pretest administration, I used the TEMA pretest as a Level 1 covariate. Including the pretest as covariate helped provide statistical precision and reduce error variance or unmodeled variability within classrooms (Konstantopoulus, 2008). As recommended by Raudenbush and Bryk (2002), the covariate was centered around the grand mean.

The second predictor variable at the student level was risk status in the beginning of kindergarten. To establish risk status, the study utilized a 20th-percentile cutoff as assessed by two curriculum-based measures (Oral Counting [OC] and Number Identification [NI]; Clarke & Shinn, 2004). Initial skill performances at or below the 20th percentile on the OC and NI pretests were raw scores of 12 and 14, respectively. Because oral counting and numerical identification are underlying components of the number sense construct (Berch, 2005; Gersten & Chard, 1999), this study considered students at or below the 20th percentiles on both the OC and NI measures as at risk for mathematics difficulties. From the analytic sample (*N* = 929), approximately 119 students (12.8%) were determined as at risk. The Risk predictor was coded 0 for students considered at low risk and 1 for students considered at risk for difficulties in mathematics. This predictor remained uncentered in the models. The predictive validity coefficients between initial skill performance on OC and NI and posttest TEMA were .61 and .67, respectively. The validity coefficient between OC and NI was .62.

Classroom-Level Predictors

The first classroom-level predictor used in the study was treatment condition. The ELM-ETKC randomly assigned classrooms to either treatment or a comparison condition. Instruction in the treatment condition consisted of the Early Learning in Mathematics curriculum. Instructional materials and teaching approaches varied across the comparison classrooms. The treatment condition predictor was coded 0 for classrooms randomly assigned to the comparison condition and 1 for classrooms randomly assigned to the treatment condition. This predictor remained uncentered in the model.

The second classroom-level predictor was round of observations. As previously noted, the ELM-ETKC project planned three different occasions of direct observation per classroom for the 2008-2009 school year. This variable, which remained uncentered, represents time or the three observation rounds. For this predictor, the first round of observations (fall administration) took on the value of one, while the second and third rounds took on the values of two and three, respectively.

The third classroom-level predictor was the mean rate of teacher models. The predictor represents the mean rate of teacher models aggregated across the three observation rounds. The fourth and fifth classroom-level predictors were the mean rates of group responses and rate of individual responses. These predictors represent the mean rate of behaviors aggregated across the three observation rounds. All three predictors were centered around their respective grand mean. Analyses used the rate-per-minute metric to better control for differences in duration of classroom observations.

Teacher Demographic Survey

The survey is a researcher-developed instrument that obtains background information of participating ELM-ETKC teachers. The 13-item survey obtains demographic information about teacher ethnicity, age, gender, teaching experience, education, and areas of specialization. The survey also elicits information about class size, number of students at risk for failure in mathematics, and previously used mathematics materials. All teachers were administered the survey at the start of the ELM-ETKC study. It is important to note that this study used the survey information for descriptive purposes only.

Classroom and Student Characteristics Questionnaire

The questionnaire is a researcher-developed instrument that obtains information from participating ELM-ETKC teachers about the features of mathematics instruction and characteristics of learners. At the math instruction level, the questionnaire elicits information about the amount of mathematics instruction provided, number of children receiving special education and English language services, primary language used during math instruction, and amount of support received from an instructional assistant. At the student characteristic level, the questionnaire elicits information about each student participating in the ELM-ETKC study. The student section comprises items related to a student's first language, attention of the student during mathematics instruction, degree of absenteeism from math instruction, and type of special education services received. The ELM-ETKC administered the questionnaire at end of the 2008-2009 school year. It is important to note that this study used the information from the questionnaire for descriptive purposes only.

Content Validity Survey

The survey is a 12-item instrument designed specifically for the dissertation. The purpose of the survey was to assess whether the CATS observation instrument demonstrated evidence of content validity. The survey comprises questions about the content relevance and representativeness of the CATS instrument. External reviewers, unaffiliated with the ELM-ETKC project, completed the 12 items via an online survey service. This dissertation used information obtained from the online survey to address the first research question.

Test of Early Mathematics Ability-Third Edition

The Test of Early Mathematics Ability-Third Edition (TEMA-3; Ginsburg & Baroody, 2007) is a norm-referenced, individually administered measure of beginning mathematical ability. The TEMA-3 assesses mathematical understanding at the formal and informal levels for children ranging in age from 3 to 8 years 11 months. The TEMA-3 addresses children's conceptual and procedural understanding of mathematics, including counting and basic calculations. The TEMA-3 reports alternate-form and testretest reliabilities of .97 and .82 to .93, respectively. For concurrent validity with other math outcome measures, the TEMA-3 reports coefficients ranging from .54 to .91. The ELM-ETKC project administered the TEMA-3 at pre- (fall) and posttest (spring) measurement periods in both treatment and comparison classrooms. For this study, the TEMA pretest score served as a covariate. TEMA posttest scores served as student outcomes. From the analytic sample, student TEMA posttest performances at the 25th and 50th percentiles consisted of raw scores of 27 and 34, respectively.

Oral Counting

Oral counting (OC) is a standardized, individually administered curriculum-based measure (Clarke & Shinn, 2004). The measure assesses a student's counting ability. For this one-minute measure, a student orally counts as high as possible without making an error. The OC discontinue rule applies after the first counting error. Previous research studies report concurrent and predictive validity correlations ranging from .46 to .72 (Clarke & Shinn, 2004; Lembke, Foegen, Whittaker, & Hampton, 2008). On the basis of the analytic sample, the association between OC and TEMA pretest scores (concurrent validity) was strong (.77). The predictive validity coefficient between OC and TEMA posttest scores was moderate to strong (.61). This study used initial performance on OC (pretest raw score) to establish a risk status cut-off.

Number Identification

Number identification (NI) is a standardized, individually administered curriculum-based measure (Clarke & Shinn, 2004). NI assesses a student's ability to read numerals. For this one-minute measure, a student orally identifies numerals between 0 and 10. The order of presentation for all numerals is random. Alternate form reliability ranged between .89 and .93 (Clarke & Shinn, 2004). Concurrent and predictive validity correlations ranged from .68 to .71 (Clarke & Shinn, 2004; Lembke et al., 2008). On the basis of the analytic sample, the association between NI and TEMA pretest scores (concurrent validity) was strong (.74). The predictive validity coefficient between NI and TEMA posttest scores was moderate (.67). Initial performance on NI (pretest raw score) helped establish a risk status cut-off.

Data Analysis

Content Validity (Research Question 1)

To assess the relevance and representativeness of the CATS tool (Messick, 1995), I completed a content-related review. The purpose of the review was to justify the use of the CATS in measuring instructional interactions during kindergarten math instruction. I anticipated this external review would help corroborate the proposed content and uses of the CATS observation instrument.

For the analysis, an online survey for external reviewers was created using SurveyMonkey.com, a free service that offers survey software. The survey consisted of

12 items and required approximately 10 to 15 minutes to complete. Each behavior represented two items. The survey also included areas for the reviewers to note suggestions for improving the observation instrument.

Six items pertained to the relevance of the behaviors captured by the observation instrument. For instance, the first two items addressed the content relevance of each teacher behavior: teacher model and academic feedback. The next four items addressed the relevance of each student behavior: group response, individual response, covert response, and student mistake. These items asked evaluators to rate the extent to which each behavior was relevant to kindergarten mathematics instruction. For these items, evaluators used a 4-point scale, ranging from *irrelevant* (1) to *highly relevant* (4). A rating of 4 represented the highest score.

For the final six items, the survey asked evaluators to rate the extent to which each behavior represented the important instructional interactions that occur between a teacher and her students during kindergarten mathematics instruction. For these items, evaluators used a 4-point scale, ranging from *not at all* (1) to *highly representative* (4). A rating of 4 represented the highest score. I considered a mean score of 3 as an average acceptable score for all items.

Management of the content analysis involved a four-step process. First, the online survey was developed through a series of initial iterations. Two members of the ELM-ETKC project provided feedback during the development process. Second, a list of 12 prominent educational researchers was compiled. Reviewers were selected based on their expertise in elementary math curricula and instruction, and their authorship of peer-reviewed publications in the field of educational research. All experts selected for the content analysis were unaffiliated with the ELM-ETKC project, the University of Oregon, and the development of the direct observation system. Third, potential reviewers were sent a recruitment email that described the purpose of the online survey and the nature of the direct observation system. The email directed the reviewers to click on the URL link http://www.surveymonkey.com/MySurveys.aspx to complete the online survey. After the initial email contact, experts did not receive a follow-up request. Fourth, approximately 30 days after the contact email and survey posting, data were collected and analyzed using the SurveyMonkey.com software.

Of the 12 experts contacted via email, 7 (or 58%) responded and agreed to complete the content analysis. All respondents identified as university faculty. For the five experts who declined to participate, none stated their reasons for not responding.

Interobserver Agreement (Research Question 2)

To address the second research question, I measured agreement among independent observers across a series of paired observations. As noted earlier, a paired observation consisted of two persons observing the same event. I anticipated observers would document minimally acceptable values of interobserver agreement given the level of observation training provided.

In all, the ELM-ETKC project checked interobserver agreement during 24% (*n* = 46) of all classroom observations. This research question utilized information collected during the 46 observation pairings. Observers completed agreement checks in both treatment and comparison classrooms across the three observation rounds. Twelve agreement checks took place in the first round along with 18 and 16 in the second and third rounds, respectively. The purpose of the pairings was to demonstrate that data obtained by the CATS instrument were consistent across the 12 members of the ELM-ETKC observation team.

To calculate observer agreement, I applied the frequency-ratio index, which estimates the overall occurrence of teacher and student behaviors (Hintze, 2005; Kennedy, 2005; Suen & Ary, 1989). As previously noted, this index takes the sum of behavior codes from two observers and divides the smaller value by the larger value. Resulting agreement values range from 0.00 to 1.00, with 1.00 being perfect agreement. A value of 0.00 indicates no agreement between two observers. By convention, a minimally acceptable standard of interobserver agreement when using this index is .80 (Horner et al., 2005; Suen & Ary, 1989). For each paired observation, the study calculated five interobserver agreements using the frequency-ratio approach: (a) agreement of all observed behaviors; (b) agreement of teacher behaviors, which collapsed teacher models and academic feedback; (c) agreement of student practice opportunities, which collapsed individual, covert and group responses; (d) agreement of teacher models; and (e) agreement of academic feedback.

Previous research studies have used the frequency-ratio approach to calculate interobserver agreement (Hart, 1983; Jason & Liotta, 1982; Murray, Hutchinson, & Bailey, 1983). The current study used this index based on the design of CATS. Recall that CATS is an event recording system and does not separate classroom observations into discrete time intervals (Kennedy, 2005). CATS requires observers to code each time a student or teacher behavior of interest occurs. When observation systems, like CATS, employ a noninterval approach, there is a chance of observers getting out of coding sequence. For example, if observer A records a teacher behavior within the first minute of the observation and observer B misses the same behavior, then subsequent codes would be scored differently by the two observers. In this instance, a missed behavior may lead to biased estimates of observer agreement, particularly with indices that require time sampling methods, such as Cohen's Kappa (Cohen, 1960) and overall agreement (Kazdin, 1982).

Although the frequency-ratio approach is not a chance-corrected form of reliability (Feuerman & Miller 2005; Suen & Ary, 1989), educational researchers recognize it as an acceptable index of interobserver agreement. For example, in a review of five interobserver agreement indices, Hintze (2005) discussed the general appeal of the smaller/larger index or frequency-ratio approach. Hintze also noted that the smaller/index "should only be used in cases where other more meaningful measures of agreement cannot be established" (p. 510). The current study is such a case.

Intercorrelations

Intercorrelations were calculated to test for multicollinearity among the observed behaviors. The calculations include intercorrelations among the rate-per-minute score for group responses, individual responses, and teacher models. Intercorrelations were estimated for the 191 observations conducted by the ELM-ETKC project.

Multilevel Models

Because of the hierarchical nature of the data, this dissertation used Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) to address the third and fourth research questions. Each of these research questions fit multilevel models (two-level) for the dependent variables, rates of observed behaviors and student posttest scores. As suggested by Raudenbush and Bryk (2002), the study employed an incremental process to develop each multilevel model. For example, to partition the variance in the dependent variable that existed between and within the Level 2 units, the process began with an unconditional model with no Level 1 or Level 2 predictors. Next, the process incorporated predictors at Level 1. Following examination of the Level 1 predictors, predictors at Level 2 were added. For the fitted multilevel model, regression parameters and variance components were examined. All multilevel models used SPSS 15.0 (SPSS, 2006) and HLM-6 (Raudenbush, Bryk, Cheong, & Congdon, 2004) software. Analyses used an alpha level of .05 as a cutoff for statistical significance.

Discriminant Validity (Research Question 3)

To examine whether the observation instrument was sensitive to instructional differences, I fit separate multilevel models (two-level) and predicted the rate-per-minute scores of three observed behaviors. The outcome variables or rate-per-minute scores for the models included the mean rate of teacher models, group responses, and individual responses. Each multilevel model nested repeated observations (i.e., Level 1) within classrooms (Level 2; Raudenbush & Bryk, 2002), using observation "ROUND" as a Level 1 predictor and treatment condition as a Level 2 predictor. Variables at each level of the model were uncentered. I present an example for all three models in the equations below.

$$\mathrm{Y}_i = \pi_{0i} + \pi_{1i}(R_i) + e_i \qquad e_{ti} \sim N(0, \sigma^2) \qquad (1.10)$$

$$\pi_{0i} = \beta_{00} + \beta_{01}(C_i) + r_{0i} \qquad r_{0i} \sim N(0, \tau^2) \qquad (1.20)$$

In Equation 1.10 or Level 1, Y_i represents an average rate-per-minute score for an observed behavior (e.g., mean rate of teacher models) for classroom *i*; while π_{0i}, the intercept parameter, is the mean rate of an observed behavior for a comparison classroom; π_{1i}, the slope parameter, is the expected change of observed behaviors for classroom *i* across the three observation time rounds (R_i: 1 = fall, 2 = winter, 3 = spring); and e_i represents a Level 1 error term. An assumption is that the error term independently and normally distributed with a mean of zero, and constant variance, σ^2 (Raudenbush & Bryk, 2002).

The classroom level of the model (Equations 1.20 and 1.30) presents the two Level 1 parameters, π_{0i} and π_{1t}, and a predictor: *C* (a dummy variable indicating random assignment of classroom condition: 1 = treatment, 0 = comparison). At the intercept parameter, π_{0j} represents the classroom average for an observed behavior; while β_{00} indicates the intercept; $\beta_{01} \cdot C$ represents the effect of condition and r_{0i} is a classroom-level error term. Analysis at Level 2 allowed for examination of condition effect on the rate of providing teacher models, group responses, and individual responses. A statistically significant variance component would indicate that rates of the observed behaviors varied across classroom conditions. For this research question, the analytic sample at Level 1 involved 191 observations and Level 2 involved 65 classrooms.

Criterion Validity (Research Question #4)

To examine the relationship between observed instructional behaviors and student math achievement, I fit a two-level model, nesting students (Level 1) within classrooms (Level 2; Raudenbush & Bryk, 2002). The model predicted covariate-adjusted, classroom-level student scores on the TEMA-3 posttest with the mean rate of observed behaviors. At Level 1, the model incorporated (a) a student-level covariate (i.e., Test of Early Mathematics Ability [TEMA] pretest) that was grand-mean centered; and (b) a predictor of risk status (0 = Low risk; 1 = At risk). The risk variable utilized initial performances (raw scores) on pretest measures of Oral Counting (OC) and Number Identification (NI). Students scoring at or below the 20th percentiles on both OC and NI were considered at risk for mathematics difficulties. Approximately 119 students (12.8%) were determined at risk in the analytic sample. The risk variable remained uncentered.

The classroom level for each model (Level 2) presents the three Level 1 parameters, B_{0j}, B_{1j}, and B_{2j}, and one Level 2 predictor. The Level 2 predictor, *ObsvBeh*, is a continuous variable indicating the average rate of observed behaviors aggregated across the three observation rounds. Three rate-per-minute scores were incorporated as the Level 2 predictors, including the rate of teacher models, the rate of group responses, and the rate of individual responses. All three Level 2 predictors were grand-mean centered. I present an example of an intercepts- and slopes-as-outcomes model in the equations below.

$$\mathrm{Y}_{ij} = B_{0j} + B_{1j}\,\mathrm{Pr}e\_TEMA_{ij} + B_2 Risk + r_{ij} \qquad r_{ij} \sim N(0,\sigma^2) \qquad (2.10)$$

$$B_{0j} = \gamma_{00} + \gamma_{01}(ObsvBeh_j) + u_{0j} \qquad u_{0j} \sim N(0,\tau^2) \qquad (2.20)$$

$$B_{1j} = \gamma_{10} + \gamma_{11}(ObsvBeh_j) + u_{1j} \qquad u_{1j} \sim N(0,\tau^2) \qquad (2.30)$$

$$B_{2j} = \gamma_{20} + \gamma_{21}(ObsvBeh_j) + u_{2j} \qquad u_{2j} \sim N(0,\tau^2) \qquad (2.40)$$

In Equation 2.10 or Level-1, Y_{ij} represents a TEMA-3 posttest score for student *i* in classroom *j*; the intercept, $B_{0\mathrm{j}}$, represents an average posttest score for a low-risk student in a classroom that provides an average rate of observed behaviors; $B_{1\mathrm{j}}$(Pre_TEMA$_{ij}$) is the covariate effect; $B_{2\mathrm{j}}$(RISK$_{ij}$) is the effect of risk status (i.e., pretest performances on both Oral Counting and Number Identification at or below the 20th percentile) on posttest scores for a student in classroom *j;* and r_{ij} represents a Level 1 error term. The error term is assumed to be independently and normally distributed with a mean of zero, and constant variance, σ^2 (Raudenbush & Bryk, 2002).

Equation 2.20 consists of the intercept parameter, γ_{0j}; the slope of observation behaviors, $\gamma_{01} ObsvBeh_j$; and u_{0j}, a classroom-level error term. Equation 2.30 represents the covariate value, γ_{10j}, for the relationship between the TEMA pretest and posttest; a cross-level interaction between TEMA pretest and the rate of observed behaviors, $\gamma_{11}ObsvBeh_j$; and u_{1j}, a classroom-level error term.

Equation 2.40, represents the relationship between RISK and TEMA posttest, γ_{20j}; a cross-level interaction between RISK and the rate of observed behaviors, $\gamma_{21}ObsvBeh_j$; and u_{2j}, a classroom-level error term. For this research question, the analytic sample at Level 1 involved 929 students and Level 2 involved 191 observations.

To estimate the proportion of variance explained between and within classrooms, I calculated an R^2 statistic for the Level 1 and Level 2 predictors. In multilevel modeling, the R^2 statistic is analogous to an eta-squared effect size from ANOVA (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). For example, at Level 2, the R^2 statistic estimates the proportion of parameter variation explained between classrooms. The R^2 statistic for the Level 1 predictors estimates the proportion of variance accounted for within classrooms. R^2 calculations were computed using the differences between a fitted model and a baseline model.

CHAPTER IV

RESULTS

The following chapter provides results of the study in five sections. The first four sections address (a) the issue of missing data, (b) descriptive statistics for student-level and classroom-level data, (c) diagnostic information related to the multilevel models, and (d) intercorrelations among the observed behaviors. The final section presents results for each research question.

Missing Student Data

This study used extant data provided by the Early Learning in Mathematics: Efficacy Trials in Kindergarten Classrooms (ELM-ETKC; Baker et al., 2008) project. Across the 2008-2009 school year, the ELM-ETKC project involved 1,495 students. Of the 1,495 students, 1,073 were assessed on the Test of Early Mathematics Ability-Third Edition (TEMA-3; Ginsburg & Baroody, 2007) at the fall (pretest) administration. ELM-ETKC assessed approximately 1,246 students in the spring (posttest). Preliminary exploration of the TEMA scores revealed that approximately 28% of the pretest data (i.e., 422 cases) and 17% of the posttest data (i.e., 249 cases) were missing.

With regard to the Oral Counting (OC; Clarke & Shinn, 2004) measure, the ELM-ETKC project assessed 1,200 students at the pretest administration. Exploration of the OC pretest data revealed 295 missing cases (20%). For the Number Identification

(NI; Clarke & Shinn, 2004) measure, the project assessed 1,164 students at the pretest administration, with 331 (22%) cases missing.

From the ELM-ETKC sample, an independent samples *t* test was conducted to determine whether significant mean differences on posttest TEMA existed between two groups of students. The first group (Posttest-only) consisted of students who participated in the posttest TEMA administration but did not participate in pretest administration of Oral Counting, Number Identification, and TEMA. The second group (Pretest-posttest) consisted of students who participated in all three pretest measures and the posttest TEMA administration. Results of the *t* test revealed a statistically reliable difference between the mean posttest TEMA score of students in the Posttest-only group ($M = 25.51$, $SD = 9.66$) and students in the Pretest-posttest group ($M = 32.13$, $SD = 9.48$), $t(1244) = .71$, $p < .001$).

To handle the large number of missing cases in the ELM-ETKC sample, this dissertation included only those students who participated in the fall and spring administrations of the outcome measure (TEMA) and the pretest administrations for both curriculum-based measures (Oral Counting and Number Identification). The analytic sample at the student level involved 929 kindergarteners. At the classroom level, the analytic sample involved 65 classrooms and 191 classroom observations.

Descriptive Statistics

Table 2 provides descriptive information, including means, standard deviations, and sample sizes, for the average rate of behaviors captured by the direct observation

TABLE 2. Descriptive Statistics Per Observation Round for Average Rate of Observed Behaviors Across Treatment Conditions

	Round 1		Round 2		Round 3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Rate of individual responses						
Treatment	0.64	0.43	0.73	0.30	0.66	0.55
Comparison	0.44	0.44	0.38	0.56	0.45	0.30
Rate of group responses						
Treatment	1.88	0.81	1.61	1.49	1.45	0.79
Comparison	0.89	0.62	1.16	0.80	0.94	0.73
Rate of teacher models						
Treatment	0.70	0.45	0.66	0.40	0.61	0.39
Comparison	0.56	0.36	0.63	0.52	0.62	0.48

Note. Total observations by round: Round-1 treatment (34), comparison (28); Round-2 treatment (35), comparison (29); Round-3 treatment (35), comparison (30)

instrument during the 191 classroom observations. Information in Table 2, reported by condition, shows distinct differences between treatment and comparison classrooms across the three observation rounds. Table 3 provides descriptive statistics for the average rate of observed behaviors aggregated across the three observation rounds. The standard deviations indicate large amounts of variability for all three rates of behavior. In fact, some classrooms were highly interactive with instruction (i.e., high rates of behaviors), whereas others demonstrated low rates of instructional interactions. This finding is noteworthy because it suggests that children may be experiencing different levels of interaction during their first year of formal instruction. For example in Table 3, a classroom of high interactions (i.e., one standard deviation above the mean)

TABLE 3. Descriptive Statistics for Average Rate of Observed Behaviors Aggregated Across Observation Rounds

Observed behavior	*M*	*SD*	Minimum	Maximum
Rate of individual responses	0.55	0.36	0.02	1.88
Rate of group responses	1.29	0.68	0.22	2.78
Rate of teacher models	0.61	0.28	0.03	1.52

Note. Total observations ($N = 191$).

demonstrates a rate of individual responses that is nearly five times greater than a classroom of low interactions (i.e., one standard deviation below the mean).

Table 4 provides descriptive statistics related to student pretest and posttest performances on the TEMA measure as well as pretest performances on the curriculum-based measures of Oral Counting and Number Identification. Descriptive statistics reported in Table 4 are presented for all participating students, regardless of treatment condition. Minimum and maximum raw scores are also reported. The information in Table 4 indicates high variability across all four measures.

Table 5 presents descriptive statistics related to the context codes captured by the observation instrument. This information includes the duration of the direct observations and the number of students observed across conditions for each observation round. Minimal differences are noted between classroom condition and across observation rounds for both context codes.

TABLE 4. Descriptive Statistics of Student Pretest and Posttest Performances on Test of Early Mathematics Ability-Third Edition (TEMA), and Pretest Performances on Oral Counting (OC) and Number Identification (NI) Curriculum-Based Measures

Measure	*M*	*SD*	Minimum	Maximum
TEMA pretest	20.29	9.58	0	56
TEMA posttest	33.12	8.99	7	70
OC pretest	26.50	20.68	0	109
NI pretest	33.36	19.65	0	108

Note. *N* = 929; OC = Oral counting; NI = Number identification; Performances below the 20th percentile on the OC and NI pretests were raw scores of 12 and 14, respectively. Performances at the 25th and 50th percentiles on posttest TEMA consisted of raw scores of 27 and 34, respectively.

TABLE 5. Descriptive Statistics for Context Codes Across Conditions Per Observation Round

	ELM			Comparison		
Code	*M*	*SD*	*n*	*M*	*SD*	*n*
Observation duration						
Round 1	37.74	9.11	34	37.00	11.55	28
Round 2	37.02	8.40	35	30.83	10.00	29
Round 3	38.20	9.12	35	34.33	10.80	30
Number of students						
Round 1	19.86	3.98	34	19.40	4.06	28
Round 2	19.00	4.32	35	19.63	3.13	29
Round 3	19.12	3.88	35	19.17	4.02	30

Note. Total observations (*N* = 191).

Diagnostics of the Multilevel Models

To determine the adequacy of the multilevel models (Luke, 2004), the study conducted three diagnostic assumption checks. First, SPSS software was used to explore (a) distributions of student performances for the TEMA pretest and posttest administrations, and (b) the pretests on Oral Counting and Number Identification (NI). Examination of the TEMA performance distributions revealed acceptable normality across the 929 student participants. For the OC and NI pretests, there appeared to be a positive skew among the distribution of scores. Because multilevel modeling is robust to violations of normality (Fitzmaurice, Laird, & Ware, 2004; Hox, 2002; Maas & Hox, 2004), the slight skewness in the OC and NI distributions was not expected to bias results of the study. Next explored was the distribution of observed behaviors across the 191 classroom observations (i.e., rate of teacher models, rate of group responses, and rate of individual responses). Examination for each observed behavior revealed acceptable normality. Finally, a test of homogeneity of Level 1 variance revealed no problems with heteroscedasticity, $\chi^2(46) = 80.70, p < .01$.

Intercorrelations

Table 6 presents correlations among the rates of observed behaviors. Correlations between the observed behavior rates ranged from .01 to .48. Findings revealed a moderate relationship ($r = .48$) between the rate of teacher models and rate of group responses.

TABLE 6. Intercorrelations Among Rate of Observed Behaviors

Codes		1	2	3
Rate of teacher models	1	1.00		
Rate of group responses	2	.48*	1.00	
Rate of individual responses	3	.01	.22*	1.00

**p* < .05.

Content Validity

An external review of the CATS instrument was conducted to assess whether the six observed behaviors provided evidence of content validity (Messick, 1995). The purpose of the review was to have experts judge the instrument's content relevance and representativeness. The target population for the survey was a group of experts with extensive knowledge in the areas of elementary mathematics curricula and instruction. Prior to contacting potential reviewers, selection criteria were established that would qualify them as experts. Reviewers considered eligible were required to meet the following criteria: (a) publication of several articles in peer-reviewed educational journals and (b) lack of affiliation with the ELM-ETKC study, the University of Oregon, and the development of the CATS instrument. This selection process gleaned twelve experts from the field. Potential reviewers received a contact email and a URL link to an online survey.

The survey was comprised of 12 items. Six contained information about content relevance and six contained information about content representativeness (Messick, 1995). Items on the survey used a 4-point scale, with 1 being the lowest and 4 the highest. To meet the average acceptable score, items had to receive a mean score of 3 or higher. When rating each item, reviewers clicked a box with the appropriate rating. The survey took approximately 10 to 15 minutes to complete.

Of the 12 participants contacted, responses were received from 7 experts (58%). Six respondents answered all items, while one respondent answered four of the 12 items. All respondents identified as university faculty.

Regarding the content relevance of the six behaviors, ratings were found acceptable. Teacher model and academic feedback both received an average rating of 4 on the content relevance items. Of the four student behaviors, student mistake and individual response received an average rating of 4. The covert response and group response behaviors met the minimum rating, earning average ratings of 3.33 (*SD* =.52) and 3.83 (*SD* =.42), respectively.

In regard to content representativeness (Messick, 1995), all six items met the minimum mean rating of 3. In fact, teacher model, academic feedback, individual response, and student mistake received average ratings of 4. Group response received a mean rating of 3.83 (*SD* =.14), while covert response earned an average of 3.67 (*SD* =.51). Collectively, reviewers rated all six behaviors above the minimum mean rating of 3.0.

Interobserver Agreement

Table 7 presents the percentage of agreements among the 46-paired observations. The ELM-ETKC project conducted 12 pairings during the first round of observations along with 18 and 16 in the second and third rounds, respectively. Twenty-six paired observations (57%) took place in treatment classrooms. One pairing, number 42 (see Table 7), captured instruction delivered in Spanish.

For every paired observation, five interobserver agreements were calculated: (a) agreement for the total number of coded behaviors; (b) agreement of teacher behaviors (i.e., teacher models and academic feedback); (c) agreement of student practice opportunities (i.e., individual, covert and group responses); (d) agreement of teacher models; and (e) agreement of academic feedback.

Across the three rounds of observations, average agreements for teacher models and academic feedback were .73 (*SD* = .19) and .77 (*SD* = .21), respectively. Observers documented the highest mean percentages with combined teacher behaviors (.84, *SD* = .13) and combined student practice opportunities (.90, *SD* = .09). Average agreement for the total number of code-able behaviors (i.e., all behaviors) was .90 (*SD* = .07).

When considering agreement across observation rounds, observers documented the strongest percentages during the third or final round. For this round, four of the five agreement calculations averaged above .80. However, average agreement of teacher models during the third round was .66. The fall and winter rounds each demonstrated

TABLE 7. Interobserver Agreements for 46 Paired (Reliability) Observations

Pairing	Condition	Round	Teacher model	Academic feedback	Teacher behaviors	Student practice	All behaviors
1	1	1	0.75	0.59	0.86	0.96	0.96
2	0	1	0.92	0.70	1.00	0.66	0.79
3	1	1	0.87	0.93	0.90	0.99	0.98
4	1	1	0.84	0.57	0.80	0.83	0.86
5	0	1	0.90	1.00	0.94	0.79	0.85
6	1	1	0.97	0.36	0.84	0.93	0.99
7	1	1	0.71	0.48	0.65	0.89	0.86
8	0	1	0.48	0.93	0.76	0.96	0.86
9	1	1	0.71	0.79	0.98	0.97	0.96
10	1	1	0.50	0.92	0.68	0.86	0.81
11	1	1	0.84	0.80	0.83	0.82	0.81
12	0	1	0.96	0.94	0.96	0.94	0.99
13	0	2	0.75	0.96	0.89	0.95	0.94
14	0	2	0.86	0.75	0.94	0.90	0.92
15	0	2	0.97	1.00	0.98	0.97	0.98
16	0	2	0.92	1.00	0.94	0.88	0.90
17	1	2	0.63	0.96	0.74	0.90	0.81
18	1	2	0.65	0.54	0.98	0.80	0.85
19	0	2	0.55	0.68	0.83	0.58	0.75
20	1	2	0.33	0.53	0.46	0.97	0.79
21	1	2	1.00	0.33	0.60	1.00	0.92
22	0	2	1.00	0.64	0.80	0.86	0.83
23	0	2	1.00	0.68	0.87	0.96	0.92
24	1	2	0.52	0.80	0.68	0.87	0.78
25	1	2	0.58	0.91	0.82	0.95	0.91
26	0	2	0.91	0.83	0.87	0.89	0.88
27	1	2	0.57	0.74	0.63	0.97	0.83
28	1	2	0.86	0.91	0.98	0.90	0.95

TABLE 7. (Continued)

Pairing	Condition	Round	Teacher model	Academic feedback	Teacher behaviors	Student practice	All behaviors
29	0	2	0.83	0.53	0.71	0.83	0.77
30	1	2	0.59	0.84	1.00	0.98	0.97
31	0	3	0.88	0.82	1.00	0.84	0.91
32	1	3	0.80	0.92	0.94	0.98	0.96
33	1	3	0.82	0.78	0.77	0.90	0.97
34	1	3	0.81	0.93	0.96	0.80	0.88
35	1	3	0.53	0.90	0.80	0.93	0.95
36	0	3	0.85	1.00	0.94	1.00	0.98
37	1	3	0.77	0.84	0.86	0.97	0.95
38	1	3	0.63	0.97	0.90	0.95	0.99
39	1	3	0.76	0.98	0.94	0.90	0.94
40	0	3	0.79	0.80	0.80	0.98	0.92
41	0	3	0.46	0.41	0.90	1.00	0.98
42	0	3	0.43	0.90	0.63	0.86	0.75
43	1	3	0.30	0.77	0.59	0.99	0.83
44	0	3	0.41	0.06	0.78	0.98	0.95
45	1	3	0.63	0.85	0.82	0.91	0.93
46	0	3	0.72	0.89	0.91	0.86	0.89

Note. $N = 46$ paired observations. Condition: treatment =1, comparison = 0. Round: 1 = fall, 2 = winter, 3 = spring.

three agreement calculations above .80. Both of these rounds had average agreements for teacher models and academic feedback ranging from .73 to .77.

Looking across treatment conditions, observers documented stronger agreement in comparison classrooms. While observers averaged an overall agreement of .90 for all behaviors captured in treatment classrooms, agreement percentages for teacher models and academic feedback were higher in comparison classrooms. For example, the average agreement for teacher models in comparison classrooms was .78 and .69 in treatment classrooms. Observers also averaged higher agreement of combined teacher behaviors in the comparison classrooms.

Sensitivity to Detect Differences Between Treatment Conditions

To assess whether the observation instrument was sensitive to treatment conditions, separate two-level, multilevel models were fit, each nesting observation occasions within classrooms. Models were tested separately for three outcome variables: rate of teacher models (MOD_RATE), rate of individual responses (IND_RATE), and rate of group responses (GRP_RATE). Each model introduced a Level 1 predictor (ROUND or observation occasion) and a Level 2 predictor (CONDITION or treatment condition). Predictor variables were introduced incrementally (i.e., ROUND then CONDITION) following the development and evaluation of the unconditional model. Variables at each level of the model remained uncentered for the analyses. All models used the full maximum likelihood (FML) method for estimation (Hox, 2002; Luke, 2004; Snijders & Bosker, 1999) and judged fixed effects against the robust standard error. Results for the unconditional models are presented first, followed by the fixed and random effects for the conditional models.

Unconditional Model

To provide a baseline model for comparison (Raudenbush & Bryk, 2002), the analysis fit separate unconditional models for the three outcomes of observed behaviors.

Unconditional models contained the Level 1 predictor, ROUND. The intercept was allowed to vary at Level 2. Table 8 presents the fixed effect for each unconditional model, while Table 9 presents the variance components. Tests for the fixed effects in the unconditional model used 64 degrees of freedom.

TABLE 8. Fixed Effects From Unconditional Model for Rate of Observed Behaviors

Fixed effect	Unstandardized coefficient	*SE*	*t*
Individual_rate			
Intercept γ_{00}	0.54	0.07	7.45*
Round γ_{10}	0.01	0.03	0.22
Group_rate			
Intercept γ_{00}	1.55	0.16	9.66*
Round γ_{00}	-0.10	0.06	-1.68
Model_rate			
Intercept γ_{00}	0.65	0.07	8.74*
Round γ_{10}	-0.01	0.03	-0.24

Note. *SE* = standard error.

$*p < .05$.

Examination of Table 8 indicates that the intercept (γ_{00}) for each model was statistically significant. Average rates of observed behaviors across classrooms ranged from .54 to 1.55 behaviors per minute. With regard to ROUND, results indicate that the

TABLE 9. Variance Components From Unconditional Models for Rates of Observed Behaviors

Random effect	Variance	*SD*	*df*	χ^2	*p*
Individual response rate					
Classroom u_{0j}	0.08	0.28	64	76.41	.138
Round u_{1j}	0.00	0.02		58.83	.500
Level-1 r_{ij}	0.12	0.34			
Group response rate					
Classroom u_{0j}	0.41	0.64	64	64.96	.443
Round u_{1j}	0.01	0.08		45.53	.500
Level-1 r_{ij}	0.67	0.82			
Model response rate					
Classroom u_{0j}	0.03	0.17	64	59.09	.500
Round u_{1j}	0.00	0.01		52.11	.500
Level-1 r_{ij}	0.16	0.40			

Note. SD = standard deviation.

Level 1 slope (γ_{10}) was not statistically significant and did not predict in the three conditional models. As can be seen in Table 9, Level 2 variance components were not statistically significant. In particular, classrooms did not significantly vary in the rate of observed behaviors and across observation rounds. Calculated intraclass correlations (ICC) showed the amount of variability in IND_RATE, GRP_RATE, and MOD_RATE that is attributable to classrooms was 44%, 26%, and 15%, respectively.

As discussed, regression coefficients and variance components for the Level 1 predictor ROUND were not statistically significant. Results of chi-square deviance tests for each outcome variable also indicated no significant differences ($p > .05$) between separate unconditional ANOVA models (Raudenbush & Bryk, 2002), without the Level 1 predictor, and conditional models that contained ROUND. For ease of interpretation, therefore, the multilevel models dropped the Level 1 predictor ROUND. Consequently, CONDITION remained as the only predictor. CONDITION, an uncentered Level 2 predictor, was coded 1 for treatment classrooms (i.e., Early Learning in Mathematics, ELM) and 0 for classrooms assigned to the comparison condition.

Results for the chi-square deviance tests, including the Level 2 predictor CONDITION, varied across the outcome variables. For IND_RATE and GRP_RATE, chi-square statistics indicated better fit of the data and the models ($p < .01$). For MOD_RATE, the deviance test was not significant ($\chi^2(2) = .57903, p > .05$). The analysis, however, retained CONDITION in the MOD_RATE model because of the predictor's fundamental importance.

Tables 10 and 11 present the fixed effects and variance components for each model, respectively. Results show statistically significant fixed effects for the average rate of behaviors in comparison classrooms for IND_RATE ($\gamma_{00} = .56, SE = .04$), $t(63) = 13.16, p < .001$; GRP_RATE ($\gamma_{00} = 1.34, SE = .07$), $t(63) = 17.78, p < .001$; and MOD_RATE ($\gamma_{00} = .63, SE = .04$), $t(63) = 17.76, p < .001$. Not surprisingly, the observation instrument did not identify significant treatment effects for rate of teacher

TABLE 10. Fixed Effects from Conditional Models for Rates of Observed Behaviors

Fixed effect	Unstandardized coefficient	*SE*	*t*
Individual response rate			
Intercept β_{00}	0.56	0.04	13.16*
Condition β_{01}	0.26	0.08	3.15*
Group response rate			
Intercept β_{00}	1.34	0.07	17.78*
Condition β_{01}	0.65	0.15	4.19*
Model response rate			
Intercept β_{00}	0.63	0.04	17.76*
Condition β_{01}	0.06	0.07	0.78

Note. *SE* = standard error.

$^*p < .05$.

models (MOD_RATE). Recall that descriptive statistics in Table 2 also indicated minimal differences when comparing the rates of teacher models across treatment conditions.

Tests of the treatment effects, however, were significant for IND_RATE and GRP_RATE. When comparing the two practice opportunities, the instrument detected stronger treatment effects for the rate of group responses, as noted by the slope coefficient (γ_{01} = .65, *SE* = .15). Thus, treatment classrooms provided 1.64 group response opportunities per minute. Though the treatment effect for the rate of individual

TABLE 11. Variance Components from Conditional Models for Rates of Observed Behaviors

Random effect	Variance	*SD*	*df*	χ^2	*P*
Individual response rate					
Classroom u_{0j}	0.07	0.28	63	187.93	< .001*
Level-1 r_{ij}	0.12	0.35			
Group response rate					
Classroom u_{0j}	0.13	0.37	63	102.06	.002*
Level-1 r_{ij}	0.69	0.83			
Model response rate					
Classroom u_{0j}	0.03	0.17	63	99.34	.003*
Level-1 r_{ij}	0.16	0.40			

Note. SD = standard deviation.

$*p < .05$.

responses was smaller (γ_{01} = .26, *SE* = .08), observers coded approximately 32% more practice opportunities in ELM classrooms than in comparison classrooms.

With regard to the variance components, significant variability across classrooms for all three models was found. For example, classrooms differed from one another in the average rate of group responses, as noted by the variance component (0.13, *SD* = .38), $\chi^2(63) = 102.05, p < .01$. The standard deviations at Level 2 for IND_RATE and MOD_RATE were less than GRP_RATE. Results also indicate that IND_RATE, GRP_RATE, and MOD_RATE reduced the estimated proportion of variance between classrooms at Level 2 by approximately .17 (R^2 = 17%), .44 (R^2 = 44%), and .02 (R^2 = 2%), respectively.

In sum, the average rate of observed instructional behaviors did not change over time. However, the rate of individual and group responses differed significantly between the treatment and control conditions. Results also show that the average rate of all behaviors significantly varied from one classroom to another.

Prediction of Student Outcomes

To test the relationship between student math achievement and the three observed behaviors, I fit student-level data and classroom observation data into a two-level model (Raudenbush & Bryk, 2002). The model nested students (Level 1) within classrooms (Level 2) and predicted student math outcomes with two Level 1 predictors (RISK, Pre_TEMA) and three Level 2 predictors (GRPRATE, TCHMRATE, INDRATE). The following section provides results for the unconditional model, the conditional model with Level 1 predictors, and the fully specified model, which was an intercepts- and slopes-as-outcomes model. All models used the full maximum likelihood (FML) method for estimation and judged fixed effects against the robust standard error (Hox, 2002; Luke, 2004; Raudenbush & Bryk, 2002; Snijders & Boskers, 1999). Also, models allowed intercepts and slopes to vary randomly. To index the proportion of variance explained at Level 1 and Level 2, I compared residual estimates between models, resulting in R-squared calculations.

Unconditional ANOVA Model With Random Effects

To provide a baseline model for comparison, the analysis fit an unconditional model with no predictors at Level 1 or Level 2 for the TEMA posttest scores. The unconditional model estimated the grand mean and served as a baseline model (Raudenbush & Bryk, 2002). Results of the model show the average posttest TEMA score for participating kindergarten students across the 65 classrooms was 32.79, with $t(63) = 60.57$, $SE = .54$, $p < .01$. Table 12 displays the random effects for the unconditional model. The estimate for the Level 2 variance was 13.98 and 67.58 for Level 1. The amount of variability that lies between and within classrooms was 17% and 83%, respectively. Thus, classrooms account for approximately 17% of the variation in posttest TEMA scores among kindergarten students.

TABLE 12. Variance Components From Unconditional Model for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores

Random effect	Variance	*SD*	*df*	χ^2	*p*	% variance explained
Classroom u_{0j}	13.98	3.69	63	247.36	< .001*	17
Level-1 r_{ij}	67.58	8.22				83

Note. SD = standard deviation.

*$p < .05$.

Conditional Model With Level 1 Predictors

Table 13 displays estimates for the addition of the two Level 1 predictors, RISK and Pre_TEMA, in the conditional model. For this model and subsequent ones, the pretest TEMA predictor was centered around its grand mean (Raudenbush & Bryk, 2002). RISK, coded 1 for at-risk students and 0 for low-risk students, remained uncentered. Results of the model indicate an average posttest performance for students considered at low risk (γ_{00}) was 33.48, with $t(63)$, $p < .01$, and a robust-based standard error of .30. Results also show that the relationship between pretest TEMA and posttest TEMA was strong and statistically significant, with a coefficient value of .67 and a robust-based standard error of .03. This suggests that for every one-point increase on pretest TEMA there is a .67 point increase on posttest TEMA. Additionally, risk status was significantly related to posttest TEMA, indicating a negative regression coefficient of -2.46. This means that students considered at-risk for math difficulties score on average two and a half points lower than their typical performing peers.

Results of the chi-square deviance test indicated a significant difference ($p < .01$) between the unconditional and conditional models, with a χ^2 statistic of 849.76 and 7 degrees of freedom. Thus, adding the Level 1 predictors Pre_TEMA and RISK provided better fit of the data and the conditional model.

Table 14 presents the variance components for the conditional model. Results indicate a reduction in the amount of variability in the outcome that is attributable to the introduction of the two level-1 predictors. Specifically, the table shows that average

TABLE 13. Fixed Effects From Conditional Model With Level 1 Predictors for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores

Fixed effect	Unstandardized coefficient	*SE*	*t*
Intercept γ_{00}	33.48	0.30	110.36*
Pretest TEMA γ_{10}	0.67	0.03	24.71*
Risk γ_{20}	-2.46	0.81	-3.03*

Note. SE = Standard error.

$^*p < .05$.

TABLE 14. Variance Components From Conditional Model With Level 1 Predictors for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores

Random effect	Variance	*SD*	*df*	χ^2	*p*
Classroom u_{0j}	3.58	1.89	46	96.90	< .001*
PreTEMA slope u_{1j}	0.02	0.14	46	81.14	.001*
Risk slope u_{2j}	16.15	4.02	46	85.91	.001*
Level-1 r_{ij}	26.63	5.16			

Note. SD = standard deviation.

$^*p < .05$.

posttest TEMA intercepts significantly varied across classrooms, with a variance component of 3.58, $p < .01$. Also, pretest TEMA and RISK slopes significantly varied across classrooms. These results indicate that average pretest TEMA performances and the number of students considered at risk for math difficulties varied across classrooms.

Comparison of variance components between the conditional model and the unconditional model showed that adding Pre_TEMA and RISK as predictors of TEMA posttest scores explained 61% of the within-classroom variance at Level 1 ($R^2 = .61$).

Intercepts- and Slopes-as-Outcomes Model

To examine whether the mean rate of observed behaviors predicted student math outcomes, a two-level intercepts- and slopes-as-outcomes model (Raudenbush & Bryk, 2002) was fit to the data, incorporating the classroom-level predictors of GRPRATE (rate of group responses), TCHMRATE (rate of teacher models), and INDRATE (rate of individual responses). Classroom-level predictors were centered around their respective grand mean. The model also incorporated the student-level predictors RISK and Pre_TEMA. RISK remained uncentered in the model, while Pre_TEMA was centered around its grand mean. Results of the chi-square deviance test for the three Level 2 predictors indicated significant fit ($\chi^2(16) = 871.63, p < .01$).

Tables 15 and 16 present the results for the fully specified model's fixed effects and random effects, respectively. Estimates for the fixed effects indicate the average TEMA posttest score (γ_{00}) for students considered at low risk was 33.46, with $t(60) = 120.90, p < .001$, and a robust-based standard error of .28. Results indicate that the fixed effects for GRPRATE and TCHMRATE were not statistically significant. The role of INDRATE in predicting student posttest scores is shown by the statistically significant

TABLE 15. Fixed Effects From the Intercepts- and Slopes-As-Outcomes Model for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores

Fixed effect	Unstandardized coefficient	*SE*	*t*
Model for classroom means			
Intercept γ_{00}	33.46	0.28	120.90*
Individual response rate γ_{01}	2.11	0.86	2.46*
Group response rate γ_{02}	0.51	0.50	1.02
Teacher model rate γ_{03}	-0.51	1.08	-0.47
Model for Pre_TEMA-math posttest slopes			
Intercept γ_{10}	0.67	0.02	28.60*
Individual response rate γ_{11}	-0.20	0.06	-3.36*
Group response rate γ_{12}	-0.60	0.04	-1.43
Teacher model rate γ_{13}	0.18	0.09	2.04*
Model for RISK-math posttest slopes			
Intercept γ_{20}	-2.37	0.73	-3.27*
Individual response rate γ_{21}	-7.87	1.72	-4.57*
Group response rate γ_{22}	-0.26	1.07	-0.25
Teacher model rate γ_{23}	4.65	2.02	2.31*

Note. *SE* = Standard error.

$*p < .05$.

regression coefficient ($\gamma_{01} = 2.11, p < .05$). Thus, classrooms that provide an additional individual response per minute increase student posttest TEMA scores by 2.11 points.

TABLE 16. Variance Components From the Intercepts- and Slopes-As-Outcomes Models for Test of Early Mathematics Ability-Third Edition (TEMA) Posttest Scores

Random effect	Variance	*df*	*SD*	*p*
Classroom u_{0j}	2.75	43	1.66	.01*
PreTEMA slope u_{1j}	0.01	43	0.10	.02*
Risk slope u_{2j}	7.19	43	2.68	.01*
Level-1 r_{ij}	26.74		5.17	

Note. *SD* =Standard deviation.

$*p < .05$.

Table 15 also shows that pretest and posttest TEMA are positively related (γ_{10} = .67, $t = 28.60$), which suggests that for every 1-point gain (raw score) on pretest there is a .67 increase (raw score) on posttest. With regard to the test for cross-level interactions between the Level 2 predictors and Pre_TEMA, two are statistically significant. The negative coefficient for INDRATE (γ_{11}) indicates that it moderates or reduces the relationship between pretest and posttest TEMA. The positive coefficient for the cross-level interaction involving TCHMRATE (γ_{13}), in contrast, appears to strengthen the relationship between pretest and posttest TEMA.

Results indicate that the relationship between RISK and posttest TEMA was negative and statistically significant, $\gamma_{20} = -2.34$, $t(62) = -3.60, p < .01$. This means that students considered at risk for failure in mathematics (i.e., those whose initial scores on Oral Counting and Number Identification were below the 20th percentiles) score on average two points lower than their typically achieving peers on posttest TEMA. Interestingly, the relationship between RISK and posttest TEMA increases as a function of INDRATE (γ_{21} = -7.87). The negative coefficient for this cross-level interaction indicates that at-risk students fall further behind their typical-performing counterparts in classrooms that provide higher rates of individual response (IR) opportunities. For example, in a classroom that provides one IR per minute above the average rate, the posttest score of students at risk would decrease by roughly eight points (i.e., γ_{00} - γ_{21} = 33.46 – 7.87 = 25.59 points).

The statistically significant cross-level interaction for γ_{23} demonstrates the role of TCHMRATE in moderating the relationship between RISK and posttest TEMA ($\gamma_{23} = 4.65$, $t(43) = 2.30$, $p < .05$. In contrast to the previous cross-level interaction (INDRATE x RISK), TCHMRATE may attenuate student RISK status in a classroom with an above average rate of teacher models. This means that for each additional teacher model per minute, the posttest TEMA scores for at-risk students increase approximately 5 points. Thus, kindergarten students struggling with mathematics may benefit more in classrooms that employ higher rates of explicit instruction.

Figure 3 presents the predicted scores for two groups of students, at-risk for math difficulties and low risk for math difficulties, in four types of classrooms. The figure shows a distinct performance gap between the two groups in the first type of classroom. Students considered at low risk outperform their at-risk peers by nearly 12 points in a classroom that provides an average rate of individual responses and teacher models. Similarly, in the second type of classroom, where there is an above average rate of individual responses and an average rate of teacher models, the performance gap appears to widen between the two groups of students. This suggests that higher rates of individual practice opportunities may be more beneficial for students at low risk for difficulties. In the third type of classroom or one that provides an above average rate of teacher models but an average rate of individual responses, the predicted performance of students at-risk for math difficulties appears to increase by approximately 2 points compared to the predicted scores in the second classroom type. This indicates that higher rates of explicit instruction may be more important for students struggling to

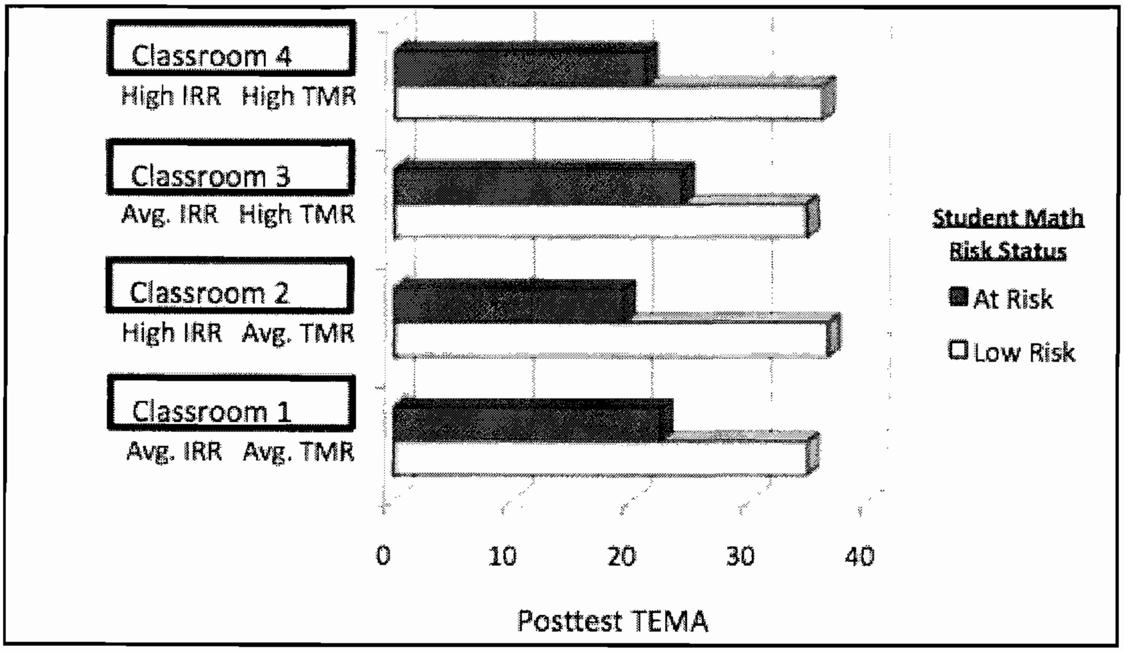


FIGURE 3. Predicted posttest TEMA scores for students with and without math difficulties in four different types of classrooms. IRR = Individual response rate; TMR = Teacher model rate; GRR = Group response rate. Assumes average rate of group responses.

learn mathematics. Finally, in the fourth type of classroom, the performance gap appears to widen between the two groups, showing a difference of 15 points in posttest TEMA scores. These predicted performances suggest that higher rates of individual practice opportunities are more important for students on track for math success at the beginning of kindergarten.

Table 16 summarizes the variance components for the fully specified model. Results indicate that TEMA posttest scores varied significantly across classrooms (variance estimate = 2.75, *SD* = 1.66), χ^2 (43) = 89.58, $p < .01$. The variances for Pre_TEMA slope and RISK slope were also statistically significant. Thus, classrooms varied from one another on slopes of pretest performances and risk status. The Level 1 residual standard deviation of 5.17 shows that students differ from one another within classrooms on Post-TEMA after taking into account the pretest covariate, risk status, and the three rates of observed classroom-level behaviors. Finally, with the addition INDRATE, GRPRATE, TCHMRATE as Level2 predictors, the estimated proportion of variance between classrooms at Level 2 in TEMA outcomes was reduced by 23% (R^2 = .23).

CHAPTER V

DISCUSSION

The purpose of this dissertation was to gather preliminary evidence in support of the Coding of Academic Teacher-Student interactions (CATS) observation instrument. The study addressed research questions related to content validity, discriminant validity, and criterion-predictive validity. Additionally, the study investigated agreement percentages between pairs of independent observers. For two of the research questions, the study employed hierarchical linear modeling (HLM) and fit separate two-level HLM models (Raudenbush & Bryk, 2002) that nested (a) observations within classrooms, and (b) student posttest math scores within classrooms.

To address the research questions, the study harvested existing data from the Early Learning in Mathematics: Efficacy Trials in Kindergarten Classrooms (ELM-ETKC) study. The data included student and classroom-level information collected in 65 kindergarten classrooms. At the student level, data included pretest and posttest scores from 929 kindergarten students on the Test of Early Mathematics Ability-Third Edition (TEMA) and two curriculum-based measures: Oral Counting and Number Identification. Information at the classroom level included observational data from 191 classroom observations.

This chapter begins with a summary table of the study's results. This is followed by the interpretation of each research question. Limitations of the study are also discussed. The chapter concludes with suggested implications for future research and practice.

Summary of Results

Table 17 summarizes the results for each research question and its corresponding hypothesis. Overall, the study found promising evidence for using CATS to measure the quantity and quality of kindergarten mathematics instruction. For example, independent observers met a minimum threshold of interobserver agreement. Also, CATS demonstrated high levels of content validity, as well as the requisite sensitivity to detect differences between treatment conditions. Finally, results from the prediction analyses showed the role of observed behaviors in predicting student math outcomes.

Evidence of Content Validity

CATS measures six behaviors related to the instructional interactions that occur between teachers and students during kindergarten mathematics instruction. An emerging, yet strong, body of scientific evidence has linked four of the behaviors to improved student outcomes. These behaviors are teacher models, academic feedback, individual response opportunities, and group response opportunities. The National Mathematics Advisory Panel (2008) strongly recommends that teachers use an explicit

TABLE 17. Summary Table of Support for Coding of Academic Teacher-Student Interactions (CATS) Observation Instrument

Research question	Hypothesis	Support for hypothesis
Content validity	CATS will demonstrate content relevance, as measured through an external content review.	Yes
Interobserver agreement	Observers will reach a minimally acceptable level of agreement across both treatment and comparison classrooms, and across observation rounds.	Yes
Discriminant validity	CATS will detect differences in the mean rate of instructional behaviors between treatment conditions.	Yes
Predictive validity	Results will show a significant relationship between student math outcomes and rates of observed behaviors, captured by CATS.	Yes

instructional approach when working with students with or at risk for math difficulties. This level of instruction entails clear and consistent teacher modeling as well as timely academic feedback. There is also consistent evidence that many students, especially those with mathematics disabilities, benefit from repeated practice opportunities (Fuchs et al., 2008; Fuchs et al., 2010; Gersten et al., 2009). These learning opportunities are most effective when distributed across groups of students and ability levels.

Because the information gathered by the CATS instrument aligns with the best evidence on effective instructional practices in early mathematics, I expected it to demonstrate acceptable evidence of content validity via an external content analysis. Findings of the external review revealed that the behaviors the CATS instrument purports to measure are relevant and representative of kindergarten mathematics instruction. In fact, reviewers rated all six behaviors above 3.0, an arbitrary acceptable score. These findings lend preliminary support for using the instrument to measure the quantity and quality of classroom instruction.

Consistency of the Observation Data

Examination for the consistency of observation data collected across 46 paired observations involved five calculations of interobserver agreement. These calculations included agreement of (a) teacher behaviors, (b) student practice opportunities, (c) teacher models, (d) academic feedback, and (e) all behaviors. Calculations used the frequency-ratio approach (Suen & Ary, 1989), and the acceptable level of observer agreement was set at 80%. In all, this research question calculated 230 estimates of interobserver agreement from the 46 paired observations.

Overall, the consistency of the data collected across independent observers and across three observation rounds was promising. Observers were able to collect reliable data across treatment conditions, strands of mathematics, and multiple time points. For example, agreement for all codable behaviors averaged 90%, with little variance between pairings (*SD* = .07). Also, observers met and maintained the minimum threshold for student practice opportunities and teacher behaviors, averaging 90% and 84% across the three observation rounds, respectively. Consistency of the data was most evident in the third or final round. It's reasonable to assume that familiarity with classroom instruction and fluency with CATS contributed to this finding. Surprisingly, observers were more consistent in comparison classrooms. Given the systematic structure of the Early Learning in Mathematics curriculum, I expected observers would reach higher agreement in treatment classrooms.

When examining the separate estimates for teacher models and academic feedback, percentages were among the lowest documented. Observers averaged 73% for teacher models and 77% for academic feedback. This finding was not surprising. Despite extensive training with the observation instrument, observers expressed difficulty in discriminating between these two behaviors during real-time observations. One explanation is that teacher models and academic feedback require a higher level of observer inference than the four student behaviors. To further complicate things, this interpretation has to occur within a matter of seconds. For example, the CATS manual defines academic feedback as taking the form of either error correction or response affirmation. Therefore, when a teacher affirms a correct student response, his/her behavior can often resemble a teacher model, as in the following example: "*Yes, six plus two equals eight.*" Observers could easily mistake this behavior in the short time they have to decide whether to code academic feedback or teacher model. Ambiguous coding situations, like the previous example, can quickly deflate an agreement estimate.

Coding these two behaviors more reliably may require an appropriate change to CATS to eliminate the response affirmation aspect from the academic feedback code. In other words, observers would only code academic feedback if it follows a student mistake. Observers, therefore, would code all demonstrations and response affirmations as teacher models. This change may help future studies in obtaining greater data consistency.

Finally, with regard to the agreement estimates that fell below the .8 cut-off, it is important to consider two points. The first is that this study reported 13 averages of interobserver agreement. In a review of seven available observation instruments, Volpe, DiPerna, Hintze, and Shapiro (2005) found one system that reported no psychometric information about interobserver agreement coefficients. A second point to consider is that the agreement calculations that fell just below the cut-off may be acceptable for classroom observation research that involves low to moderate inference instruments (Baker, Chard, Ketterlin-Geller, Apichatabutra, & Doabler, 2009). Previous research involving direct observation has reported agreement coefficients below traditional standards (Baker et al., 2006; Gersten, Baker, et al., 2005; Jackson & Neel, 2006; Nougaret, Scruggs, & Mastropieri, 2005; Pianta, Belsksy, Vandergrift, Houts, & Morrison, 2008; Stanovich & Jordan, 1998; Stuhlman & Pianta, 2009; Sutherland, Alder, & Gunter, 2003; Volpe, McConaughy, & Hintze, 2009; Wilson, Pianta, & Stuhlman, 2007). It's possible these studies and the current one were "willing to sacrifice some degree of reliability in exchange for an increase in some aspect of validity" (Brennan, 1998, p. 6).

Discriminant Validity

The test for instrument sensitivity did not reveal statistically significant differences between classroom conditions for the mean rate of teacher models. This finding was surprising given that the Early Learning in Mathematics (ELM) curriculum contains scripted dialogue to assist teachers in model demonstrations. While observers were unable to reference all materials used in comparison classrooms, a reasonable assumption is that these instructional tools do not contain explicit opportunities for modeling math concepts and skills. Recent curricular reviews indicate that many of the mathematics programs used in today's classrooms contain weaknesses in their instructional organization. For example, in a review of three U.S. market-leading programs, Doabler, Fien, Nelson-Walker, and Baker (2010) found few opportunities for explicit instruction and teacher demonstrations.

Another assumption as to why the observation instrument did not detect differences between conditions for the rate of teacher models is that CATS does not discriminate between high-quality teacher models and low-quality ones. Thus, observers code both simple teacher models and in-depth ones in the same fashion. For example, observers would code a teacher model if a teacher were to hold up a shape and simply state its name during a geometric activity. In this same context, observers would also code a teacher model if a teacher were to present the same shape, state its name, and provide details about the shape's attributes. This latter example is more representative of the ELM program's curricular design. I argue there are distinct differences between these two examples and that students, especially those at risk for failure in math, reap greater benefit from the more detailed demonstration. One caveat to this recommendation is that too many teacher models or lengthy ones may preclude the opportunity for student practice.

For the remaining two outcomes, the average rate of group responses and rate of individual responses, the instrument detected differences in instruction. Observers coded nearly twice as many group and individual responses in treatment classrooms compared to comparison classrooms. This finding is noteworthy because it corroborates the ELM's curricular intent to provide frequent teacher-student interactions during math instruction and the instrument's capacity to measure differences across treatment conditions.

Prediction of Student Outcomes

Results from the multilevel model that predicted student math outcomes with pretest TEMA scores, student risk status, and rates of observed behaviors were both encouraging and surprising—encouraging because the effect of individual response was positive and strong ($\gamma_{01} = 2.11, p < .05$), surprising, however, because the fixed effects for the rate of group responses and rate of teacher models were not statistically significant. One plausible explanation for this latter finding is the lack of statistical power. Because the number of units at each level of a hierarchical model affects statistical power, with the number of Level 2 units in a two-level model having the greater impact (Konstantopoulus, 2008; Schochet, 2008), the 65 classrooms in this analytic sample may have been insufficient to detect significant effects for group responses and teacher models.

Interestingly, results from the pretest-posttest slopes-as-outcomes model showed contradictory findings for the cross-level interactions of teacher models and individual responses. Whereas the rate of individual responses reduced the relationship between pretest and posttest TEMA, rate of teacher models appeared to strengthen it. This suggests that on-track students at the start of kindergarten require lower rates of teacher demonstrations. Perhaps students considered at low risk require less teacher demonstration and more independent practice opportunities.

The results of the risk-posttest slopes-as-outcomes model also showed a positive cross-level interaction involving the rate of teacher models. However, in contrast to the previous slopes-as-outcomes model, teacher models reduced the relationship between a Level 1 predictor—in this case, student risk status—and TEMA posttest scores. The positive coefficient indicates how a highly interactive classroom or one that provides an above average rate of teacher models moderates the relationship between risk and math outcomes (Baron & Kenny, 1986). This suggests classrooms that employ explicit instruction may help negate the effect of being at risk for math difficulties. Figure 3 graphically depicts this cross-level interaction. The figure shows that at-risk students closed the performance gap between themselves and their on-track peers when placed in classrooms that provided teacher models at a rate above the average (i.e., approximately one model per minute).

In many ways, this finding corroborates the recommendations of expert panels (e.g., Gersten et al., 2009; NMAP, 2008) and the findings of previous research on effective math interventions for promoting mathematics proficiency for at-risk learners (e.g., Baker et al., 2002; Gersten et al., 2008; Jayanthi et al., 2008). According to this body of literature, explicit methods of instruction, such as clear explanations and demonstrations of math concepts and skills, are effective for students with or at risk for math difficulties. While this form of instructional approach should not dominate math instruction, it should, however, be a staple of classroom practice for promoting fundamental understanding of basic skills and concepts.

Although not in the expected direction, results of the slopes-as-outcomes model also showed a statistically significant negative coefficient for the cross-level interaction between risk and the rate of individual responses. At-risk students scored nearly eight points lower in classrooms with above average rates of individual practice opportunities. This result, however, should not imply that all individual practice opportunities are deleterious for at-risk students. Rather, it raises two interesting points. First, as previously discussed, CATS documents classroom-level behaviors. Moreover, the tool does not discriminate responses directed to specific subgroups of students (e.g., at risk vs. low risk). This result, therefore, may suggest that classrooms with above average rates of individual responses fail to evenly distribute these types of practice opportunities among their students. In other words, teachers may overlook struggling learners. The second point involves the level of difficulty for these individual responses. It is reasonable to assume that some practice opportunities are too difficult for at-risk students, especially when explicit teacher models and structured demonstrations do not precede the opportunities.

Limitations

This study was limited by several factors. First, the study's fourth research question examined the relationship between the rates of observed behaviors and student math outcomes. Results show preliminary support for one of the observed behaviors, rate of individual responses, to predict student math performance. However, the study did not reveal evidence for the rate of group responses and teacher models. A larger number of units at Level 1 (i.e., students) might have provided greater statistical power and possibly resulted in different findings. Recall that the analysis excluded over 450 students from the analytic sample because of missing pretest data. Given the large number of missing cases, future analyses involving this data set should use statistical procedures, such as Empirical Bayes estimation, to better address the missing pretest performances.

A second limitation was the statistical mean differences in posttest TEMA scores found between the groups of students with pretest data and those without. In the full Early Learning in Mathematics: Efficacy Trials in Kindergarten Classrooms (ELM-ETKC) sample, students who did not participate in the pretest administrations of TEMA, Oral Counting, and Number Identification scored, on average, roughly seven points lower on posttest TEMA than their classmates who did participate. In fact, students with pretest and posttest data scored just below the 50th percentile on posttest TEMA, whereas the posttest-only group scored below the 25th percentile. This result suggests that the analytic sample was missing a large number of at risk students. Plausible assumptions are that family mobility, absenteeism, or selection bias attributed to the missing values (Shadish et al., 2002).

A third limitation is the index used to calculate interobserver agreement. Although researchers commonly use the smaller/larger index (Hintze, 2005; Kennedy, 2005; Suen & Ary, 1989), this type of estimate is not a chance-corrected measure of observer agreement. As such, the smaller/larger index may fail to provide an accurate picture of whether observers saw the same behavior occur or not occur. More meaningful measures such as intraclass correlation coefficients (ICC; Hintze, 2005; Rimm-Kaufman, Kirby, Grimm, Nathanson, & Brock, 2009; Smolkowski & Gunn, 2010) and generalizability theory (Hintze & Matthews, 2004; Volpe et al., 2009) may provide more reliable estimates of the observation data.

A fourth limitation is the potential for missing a qualitative aspect of classroom instruction and student learning. Because CATS uses a frequency count approach, it's possible that the tool misses a qualitative side of classroom instruction that rating systems, open-ended notes, and student interviews often capture. A fifth limitation is the variable of academic feedback. Because academic feedback serves two purposes, capturing corrective and affirmative feedback, it is possible this dual definition attributed to the low interobserver agreements among teacher models and academic feedback. A reasonable fix would be to split the definition into two separate codes. Thus, the tool would comprise seven instructional codes. Recognizing this limitation and the others noted above, readers should interpret findings from this study with caution.

Because this study took place in the context of a larger efficacy trial, researchers are addressing the limitations noted above. For example, the efficacy trial is using additional estimates to calculate interobserver reliability. Moreover, the efficacy trial is combining student and classroom-level samples across years to ensure adequate sample sizes. Future publications will reflect these changes.

Implications for Practice

Although the purpose of this study was to validate a direct observation instrument, it has implications for improving the instructional quality of kindergarten mathematics instruction. These implications directly relate to teacher education programs and in-service professional development. For example, it seems appropriate that all teachers be trained how to implement an explicit instructional approach when working with kindergarten students struggling to learn mathematics. The CATS instrument captured teacher models that were simple in structure such as brief math definitions and overt demonstrations for how to count sets of blocks. Therefore, aspiring teachers and practicing ones could easily receive training for using these kinds of explicit instructional techniques.

Additionally, teachers could receive instructional tips for differentiating practice opportunities via ongoing classroom support (e.g., expert coaching). Within this analytic sample, for instance, classrooms provided an average of .55 individual response opportunities per minute. Thus, in a 40-minute instructional time period, classrooms offered approximately 20 individual response opportunities. This results in about one individual turn per student in a classroom of 20 kindergarteners. It is my opinion that one individual turn per student is insufficient.

Implications for Research

In a previous research study, Smolkowski and Gunn (2010) used a frequency count observational system, like the CATS instrument, to document the instructional interactions that occur during first-grade reading instruction. Results from the Smolkowski and Gunn study showed student practice opportunities accounted for a significant and meaningful amount of variation in student reading-performance gains. Outside of this recent study, few studies have used a frequency count approach to measure the quality of instruction in general education elementary-level classrooms (Sutherland & Wehby, 2001). Because a successful start in the early grades is critical for subsequent achievement growth (Bodovski & Farkas, 2007; Chard et al., 2008; Morgan et al., 2009; West et al., 2001), there is need to identify and measure the effective teaching practices that facilitate learning during this critical period of children's early education.

CATS provides a systematic approach for documenting the important variables of kindergarten mathematics instruction. Moreover, findings from this study indicate that the effect of these instructional variables differs for different groups of students. Whereas increased practice opportunities appeared more important for students at low risk, higher rates of teacher models or explicit instruction seemed more likely to benefit students at risk for math difficulties. Future studies should further investigate these interactions between instruction variables and child characteristics. Additionally, studies should investigate conditional probabilities of behavior occurrences (e.g., student mistakes followed by academic feedback) and strings of instructional sequences (e.g., teacher models followed by group and individual practice opportunities). Theses proposed analyses might help reveal additional ingredients of effective classroom instruction.

Currently, observers are using CATS in instructional contexts outside of beginning mathematics, including early literacy and small-group reading instruction for Spanish-speaking English learners. Results from this dissertation as well as these ongoing studies may help future intervention studies make better use of direct observation of classroom instruction. If a goal of educational research is to identify the potential mediating variables that influence student learning, future studies will need a reliable and valid observation system.

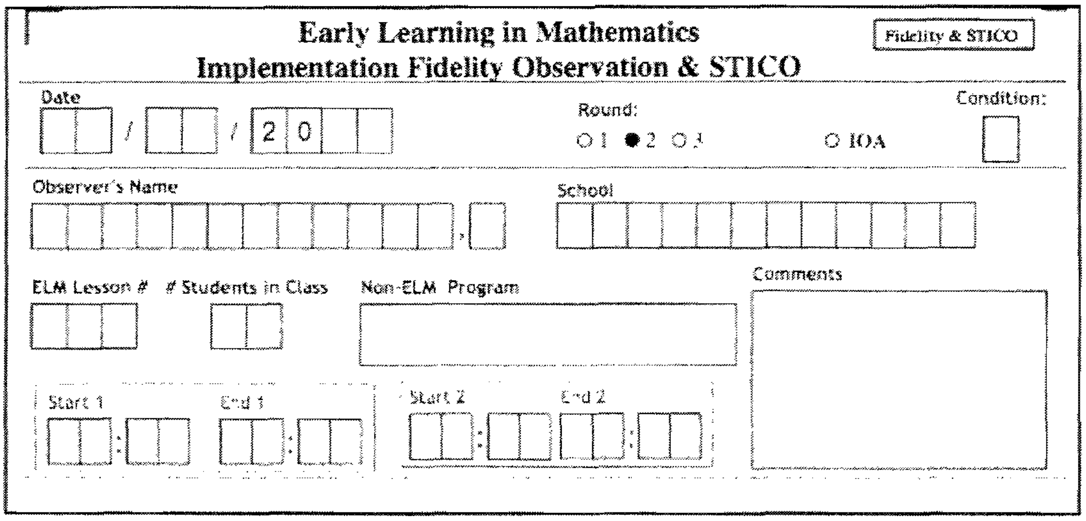
Conclusion

The recent advent of Response to Intervention calls for the delivery of high-quality instruction to occur in general education and special education classrooms. With this need, it is necessary to establish reliable and valid tools for determining whether students are receiving effective instruction in beginning mathematics. Results of this dissertation indicate that the CATS tool could potentially yield this kind of information. In particular, findings of the study provide strong, yet preliminary evidence for the observation system's capacity to capture instructional behaviors predictive of student math outcomes.

APPENDIX

CODING OF ACADEMIC TEACHER-STUDENT INTERACTIONS

DIRECT OBSERVATION INSTRUMENT



Early Learning in Mathematics
Implementation Fidelity Observation & STICO

Fidelity & STICO

Date
□□ / □□ / 2 0 □□

Round:
○ 1 ● 2 ○ 3 ○ IOA

Condition:
□

Observer's Name
□□□□□□□□□□□□, □

School
□□□□□□□□□□□□

ELM Lesson #
□□□

Students in Class
□□

Non-ELM Program
□

Comments
□

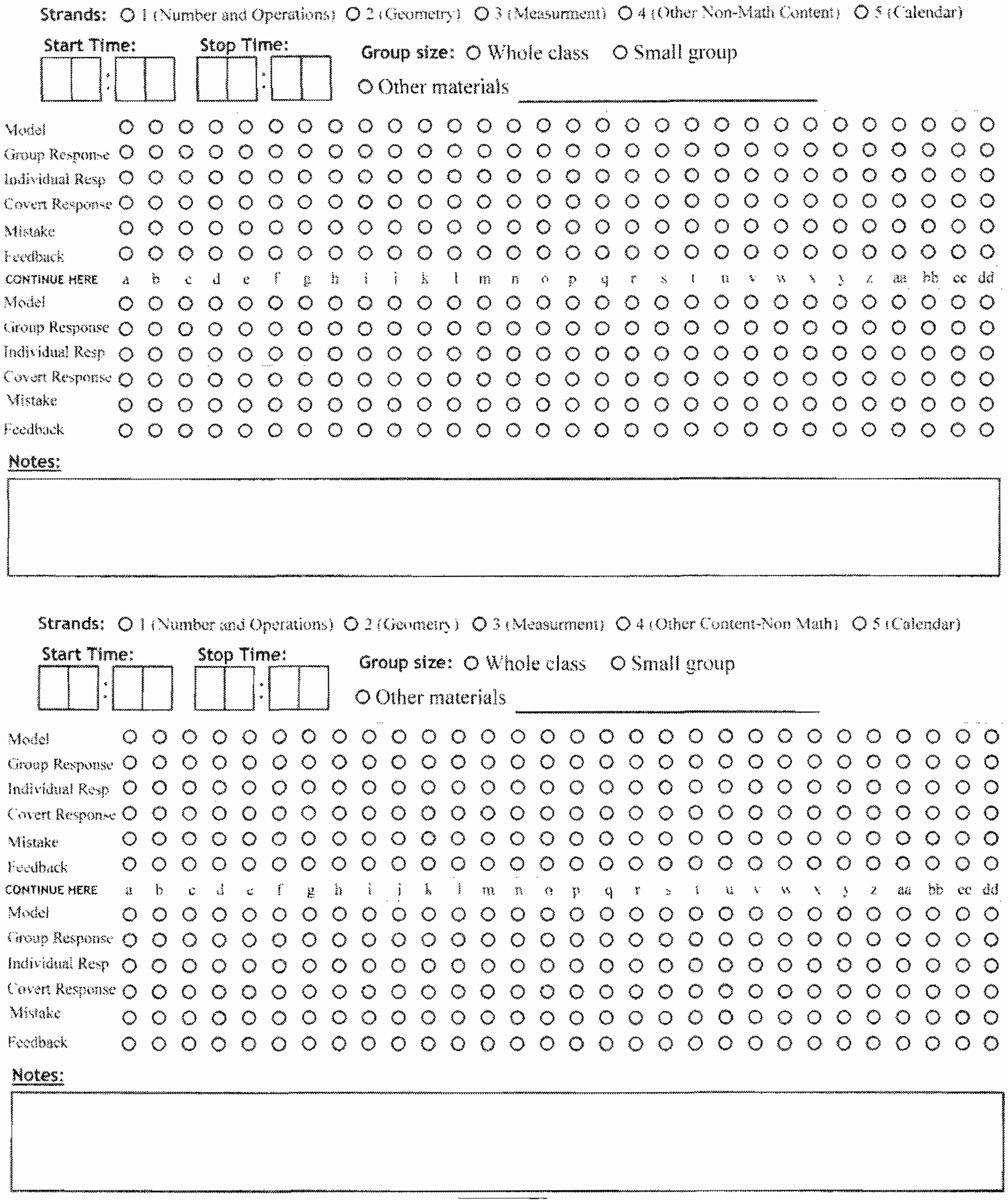
Start 1
□□:□□

End 1
□□:□□

Start 2
□□:□□

End 2
□□:□□

FIGURE 4. Cover sheet for the Coding of Academic Teacher-Student Interactions instrument.



Strands: ○ 1 (Number and Operations) ○ 2 (Geometry) ○ 3 (Measurment) ○ 4 (Other Non-Math Content) ○ 5 (Calendar)

Start Time: □□:□□ **Stop Time:** □□:□□ **Group size:** ○ Whole class ○ Small group

○ Other materials ____________

Model	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Group Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Individual Resp	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Covert Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mistake	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Feedback	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
CONTINUE HERE	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	bb	cc	dd
Model	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Group Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Individual Resp	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Covert Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mistake	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Feedback	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Notes:

Strands: ○ 1 (Number and Operations) ○ 2 (Geometry) ○ 3 (Measurment) ○ 4 (Other Content-Non Math) ○ 5 (Calendar)

Start Time: □□:□□ **Stop Time:** □□:□□ **Group size:** ○ Whole class ○ Small group

○ Other materials ____________

Model	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Group Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Individual Resp	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Covert Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mistake	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Feedback	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
CONTINUE HERE	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	bb	cc	dd
Model	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Group Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Individual Resp	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Covert Response	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mistake	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Feedback	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Notes:

FIGURE 5. Instructional interaction codes captured by the Coding of Academic Teacher-Student Interactions instrument.

REFERENCES

Anuola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96*, 699–713.

August, D., Branum-Martin, L., Cardenas-Hagan, E., & Francis, D. J. (2009). The impact of an instructional intervention on the science and language learning of middle grade English language learners. *Journal of Research on Educational Effectiveness, 2,* 345–376.

Baker, S. K., Chard, D., Clarke, B. S., Smolkowski, K., & Fien, H. (2008). *Early learning in mathematics: Efficacy trials in kindergarten classrooms* (Grant No. RA305A0814). Washington, DC: U.S. Department of Education.

Baker, S. K., Chard, D. J., Ketterlin–Geller, L. R., Apichatabutra, C., & Doabler, C. (2009). Teaching writing to at-risk students: The quality of evidence for self-regulated strategy development. *Exceptional Children, 75,* 303–318.

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first–grade English learners: Validation of an observation instrument. *Elementary School Journal, 107*, 199–219.

Baker, S. K., Gersten, R., & Lee, D. S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal, 103,* 51–73.

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *Elementary School Journal, 105,* 3–10.

Baron R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1773–1182.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: What ever happened to actual behavior? *Perspectives on Psychological Science, 2,* 396–403.

Becker, W. C., Engelmann, S., Carnine, D., & Rhine, R. (1981). The direct instruction model. In R. Rhine (Ed.), *Encouraging change in America's schools. A decade of experimentation.* New York, NY: Academic Press.

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38,* 333–339.

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *Elementary School Journal, 108,* 1157–130.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*, 5–9.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*,(4), 9–12, 27.

Brophy, J. (1999). *Teaching* (Education Practices Series No. 1). Geneva, Switzerland: International Bureau of Education. Retrieved October 18, 2008, from http://www.ibe.unesco.org

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

Bryant, B. R., Bryant, D. P., Gersten, R. M., Scammacca, N. N., Funk, C., Winter, A., et al. (2008). The effects of tier 2 intervention on the mathematics performance of first-grade students who are at risk for mathematics difficulties. *Learning Disability Quarterly, 31,* 47–63.

Carnine, D. W. (1997). Instructional design in mathematics for students with learning disabilities. *Journal of Learning Disabilities, 30*, 130–141.

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

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Carroll, J. B. (1989, January-February). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher,* 26–31.

Chard, D. J., Baker, S. K., Clarke, B., Jungjohann, K., Davis, K., & Smolkowski, K. (2008). Preventing early mathematics difficulties: The feasibility of a rigorous kindergarten mathematics curriculum. *Learning Disability Quarterly, 31*, 11–20.

Chard, D. J., & Jungjohann, K. (2006). *Scaffolding instruction for success in mathematics learning, intersection: Mathematics education sharing common grounds.* Houston, TX: Exxon-Mobil Foundation.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33,* 234–248.

Clements, D. H. (2004). Major themes and recommendations. In D. H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standard for early childhood mathematics education* (pp. 7–77). Mahwah, NJ: Lawrence Erlbaum.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45,* 443-494.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119–142.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology, National Academy of Engineering, Institute of Medicine. (2007). What actions should America take in K-12 science and mathematics education to remain prosperous in the 21^{st} century? In *Rising above the gathering storm: Energizing and employing America for a brighter economic future* (pp. 112–135). Washington, DC: National Academies Press.

Darch, C., Carnine, D., & Gersten, R. (1984). Explicit instruction in mathematics problem solving. *Journal of Educational Research, 77,* 351–359.

Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade* (NCES 2002–125). Washington, DC: National Center for Education Statistics.

Doabler, C. T., Fien, H., Nelson-Walker, N., & Baker, S. (2010). *Evaluating the instructional design elements of elementary mathematics programs.* Manuscript submitted for publication.

Engelmann, S., & Carnine, D. (1982). *Theory of instruction: Principles and applications*. Eugene, OR: ADI Press.

Ferrini-Mundy, J., & Schmidt, W. H. (2005). International comparative studies in mathematics education: Opportunities for collaboration and challenges for researchers. *Journal for Research in Mathematics Education 36,* 164–174.

Feuerman, M., & Miller, A. R. (2005). The kappa statistic as a function of sensitivity and specificity. *International Journal of Mathematical Education in Science and Technology, 36*, 517–525.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619.

Freidman, T. L. (2009, January 11). Tax cuts for teachers. *New York Times*, p. WK10.

Fuchs, L. S., Fuchs, D., & Prentice, K. (2004). Responsiveness to mathematical problem solving instruction: Comparing students at risk of mathematics disability with and without risk of reading disability. *Journal of Learning Disabilities, 37,* 293–306.

Fuchs, L. S., Powell, S. R., Hamlett, C. L., Fuchs, D., Cirino, P. T., & Fletcher, J. M. (2008). Remediating computational deficits at third grade: A randomized field trial. *Journal of Research on Educational Effectiveness, 1*(1), 2–32.

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Fuchs, D., Hamlett, C. L., Cirino, P. T., & Fletcher, J. M. (2010). A framework for remediating number combination deficits is proposed that incorporates three approaches to remediation and a two-stage system of remediation. *Exceptional Children, 76,* 135–156.

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345–362.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37,* 4–15.

Gerleman, S. L. (1987). An observational study of small-group instruction in fourth-grade mathematics classrooms. *Elementary School Journal, 88,* 3–18.

Gersten, R. (1985). Direct instruction with special education students: A review of evaluation research. *Journal of Special Education, 19*, 41–58.

Gersten, R., Baker, S. K., Haager, D., & Graves, A. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial and Special Education, 24,* 197–214.

Gersten, R., Baker, S., Pugach, M., Scanlon, D., & Chard, D. J. (2001). Contemporary research on special education teaching. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 695–722). Washington, DC: American Psychological Association.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practicguides/

Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33,* 18–28.

Gersten, R., Chard, D. J., Jaynthi, M., Baker, S., Morphy, P., & Flojo, J. (2008). Mathematics instruction for students with learning disabilities or difficulty learning mathematics: A synthesis of the intervention research. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics disabilities, *Journal of Learning Disabilities, 38,* 293–304.

Ginsburg, H. P., & Baroody, A. J. (2007). *Test of Early Mathematics Ability (Third Edition)*. Austin, TX: PRO-ED.

Glenn Commission. (2000). *Before it's too late: A report to the nation from the National Commission on Mathematics and Science Teaching for the 21st Century*. Washington, DC: U.S. Department of Education.

Goldin, C., & Katz, L. F. (2008). *The race between education and technology.* Cambridge, MA: Harvard Press.

Gonzales, P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the trends in international mathematics and science study (TIMSS) 2003* (NCES 2005–005). Washington, DC: U.S. Government Printing Office. Retrieved June 1, 2007 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005005

Good, T. L., Grouws, D. A. (1979). The Missouri mathematics effectiveness project: An experimental study in fourth grade classrooms. *Journal of Educational Psychology, 71,* 355–362.

Guarino, C. M., Hamilton, L. S., Lockwood, J. R., & Rathbun, A. H. (2006). *Teacher qualifications, instructional practices, and reading and mathematics gains of kindergartners* (NCES 2006-031). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Haas, M. (2005). Teaching methods for secondary algebra: A meta-analysis of findings. *NASSP Bulletin, 89,* 24–46.

Harniss, M. K., Carnine, D. W., Silbert, J., & Dixon, R. C. (2007). Effective strategies for teaching mathematics. In M. D. Coyne, E. J. Kame'enui, & D. W. Carnine (Eds.), *Effective teaching strategies that accommodate diverse learners* (pp. 139–170). Upper Saddle River, NJ: Pearson.

Hart, B. (1983). Assessing spontaneous speech. *Behavioral Assessment, 5,* 71–82.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning: Vol. 1. A project of the National Council Of Teachers Of Mathematics* (pp. 371–404). Charlotte, NC: Information Age Publishing.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34,* 507–519.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review, 33*, 258–270.

Hoge, R. D. (1985). The validity of direct observation measures of pupil classroom behavior. *Review of Educational Research, 55,* 469–483.

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

Hox, J. J. (2002). *Multivariate analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hudson, P., & Miller, S. P. (2006). *Designing and implementing mathematics instruction for students with diverse learning needs.* Boston: Pearson Education.

Jackson, H. G., & Neel, R. S. (2006). Observing mathematics: Do students with EBD have access to standards-based mathematics instruction? *Education and Treatment of Children, 29,* 593–614.

Jason, L. A., & Liotta, R. F., (1982). Reduction in cigarette smoking in a university cafeteria. *Journal of Applied Behavioral Analysis, 15*, 573–577.

Jayanthi, M., Gersten, R., & Baker, S. (2008). *Mathematics instruction for students with learning disabilities or difficulty learning mathematics: A guide for teachers.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Kame'enui, E. J., & Simmons, D. C., (1990). *Designing instructional strategies: The prevention of academic learning problems*. Columbus, OH: Merrill.

Kazdin, A. E. (1982). *Single–case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

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

Ketterlin-Geller, L. Jungjohann, K., Chard, D. J., & Baker, S. (2007). From arithmetic to algebra. Teachers can helps students make the transition by developing their algebraic thinking early on. *Educational Leadership, 65*(3), 66–71.

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. *Journal of Research on Educational Effectiveness, 1*, 155–178.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*, 265–288.

Kreft, I., & De Leeuw, J. (2006). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children and special education needs. *Remedial and Special Education, 24,* 97–116.

Lembke, E. S., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention, 33*, 206–214.

Levy, F., & Murnane, R. J. (2004). *The new division of labor: How computers are creating the next job market*. New York, NY: Russell Sage Foundation.

Luke, D. A. (2004). *Multilevel modeling.* Thousand Oaks, CA: Sage.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Lawrence Erlbaum.

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

Mason, D. A., & Good, T. (1996). Mathematics instruction in combination and single-grade classes: An exploratory investigation. *Teachers College Record, 98,* 236–265.

McConaughy, S. H., Ivanova, M. Y., Antshel, K., Eiraldi, R. B., & Dumenci, L. (2009). Standardized observational assessment of attention deficit hyperactivity disorder control and predominantly inattentive subtypes II. Classroom observations. *School Psychology Review, 38,* 362–381.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247–328). Chicago, IL: Rand McNally.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Miller, S. P., & Hudson, P. J. (2007). Using evidence-based practices to build mathematics competence related to conceptual, procedural, and declarative knowledge. *Learning Disabilities Research and Practice, 22*, 47–57.

Morgan, P. L., Farkas, G., & Wu, O. (2009). Five-year growth trajectories of kindergarten children with learning disabilities in mathematics. *Journal of Learning Disabilities, 42*, 306–321.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy.* New York, NY: The Free Press.

Murray, H. A., Hutchinson, J. M., & Bailey, J. S. (1983). Behavioral school psychology goes outdoors: The effect of organized games and playground aggression. *Journal of Applied Behavioral Analysis, 16*, 29–35.

National Center for Education Statistics. (2007). *NAEP 2007 mathematics: Report card for the nation and the states.* Washington, D. C.: U. S. Department of Education, Office of Educational Research and Improvement.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *The final report of the national mathematics advisory panel.* Retrieved from http://www.ed.gov/about/bdscomm/list/mathpanel/pre-report.pdf

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* J. Kilpatrick, J. Safford, & B. Findell (Eds.), Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nougaret, A. E., Scruggs, T. E., & Mastropieri, M. A. (2005). Does teacher education produce better special education teachers? *Exceptional Children, 71,* 217–229.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Evaluation and Policy Analysis, 30,* 111–140.

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice, 26*(4), 2–10.

Pianta, R. C. (2007). Preschool is school, sometimes: Making early childhood education matter. *Education Next, 7*(1), 44–49. Retrieved from http://educationnext.org/files/ednext_20071_44.pdf

Pianta, R. C., Belsksy, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45,* 365–397.

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

Pro-Ed. (2007). *Test of early mathematics ability, third edition.* Austin, TX: Author.

Rathbun, A., & West, J. (2004). *From kindergarten through third grade: Children's beginning school experiences* (NCES 2004–007). Washington, DC: U.S. Government Printing Office.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Chicago, IL: Scientific Software International.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1,* 138–154.

Rimm–Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L.L. (2009). The contributions of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 29,* 958–972.

Rosenshine, B. (1979). Content, time, and direct instruction. In J. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 28–56). Berkeley, CA: McCutchan.

Rosenshine, B. (1983). Teaching functions in instructional programs. *The Elementary School Journal, 83,* 335–351.

Rosenshine, B. (1997). Advances in research in instruction. In J. W. Lloyd, E. J. Kame'enui, & D. J. Chard (Eds.), *Issues in educating students with disabilities* (pp. 197–220). Mahwah, NJ: :Lawrence Erlbaum.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.), *Second Handbook of Research on Teaching* (pp. 122–183). Chicago, IL: Rand McNally.

Rosenshine, B., & Stevens, R. (1984). Classroom instruction in reading. In D. Pearson (Ed.), *Handbook of research on reading* (pp. 745-798). New York, NY: Longman.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher, 38*, 120–131.

Rowan, B., Correnti, R. & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospectus study of elementary schools. *Teachers College Record, 104,* 1525–1567.

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakely, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness, 1*, 89–119.

Schochet, P. Z. (2008). *Technical methods report: Statistical power for regression discontinuity designs in education evaluations* (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner (Ed.), *Handbook of educational psychology* (pp. 726–764). New York, NY: Macmillan.

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

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50–91). New York, NY: Macmillan.

Shoukri, M. M., Asyali, M. H., & Walter, S. D. (2003). Issues of cost and efficiency in the design of reliability studies. *Biometrics, 59,* 1107–1112.

Simmons, D., & Kame'enui, E. J. (1996). A focus on curriculum design: When children fail. *Focus on Exceptional Children, 28*(7), 1–16.

Simmons, D., Kame'enui, E. J, & Chard, D. J. (1998). General education teachers' assumptions about learning and students with learning disabilities: Design-of-instruction analysis. *Learning Disability Quarterly, 21,* 1–21.

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

Smolkowski, K., & Gunn, B. (2010). *The reliability and validity of student-teacher interaction and context observations collected during kindergarten reading instruction.* Manuscript submitted for publication.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snijders, T., & Bosker, R. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage.

Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science*, *7*(1), 43–56.

SPSS. (2006). SPSS 15.0 brief guide. [Computer software and manual]. Retrieved from http://www.spsss.com

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.

Stanovich, P. J., & Jordan, A. (1998). Canadian teachers' and principals' beliefs about inclusive education as predictors of effective teaching in heterogeneous classrooms. *Elementary School Journal, 98*, 221–238.

Starkey, P., & Klein, A. (2000). Fostering parental support for children's mathematical development: An intervention with Head Start families. *Early Education and Development, 11,* 659–680.

Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *Elementary School Journal, 109,* 323–342.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum.

Sutherland, K. S., Alder, N., & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with EBD. *Journal of Emotional and Behavioral Disorders, 111,* 239–248.

Sutherland, K. S., & Wehby, J. H. (2001). Exploring the relation between increased opportunities to respond to academic requests and behavioral outcomes of students with emotional and behavioral disorders: A review. *Remedial and Special Education, 22,* 113–121.

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

Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research, 76,* 249–274.

Tournaki, N. (2003). The differential effects of teaching addition through strategy instruction versus drill and practice to students with and without learning disabilities. *Journal of Learning Disabilities, 36,* 449–458.

Van de Walle, J. A. (2001). *Elementary and middle school mathematics: Teaching developmentally*. New York, NY: Addison Wesley Longman.

Vaughn, S. R., & Briggs, K. L. (Eds.). (2003). *Reading in the classroom: Systems for observing teaching and learning*. Baltimore, MD: Paul H. Brookes.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34,* 454–474.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problems and on-task scores from the direct observation form. *School Psychology Review, 38,* 382–401.

West, J., Denton, K., & Germino-Hausken, E. (2001). *America's kindergartners: Findings from the early childhood longitudinal study, kindergarten class of 1998-99* (NCES 2001-070R). Washington, DC: Government Printing Office.

White, W. A. T. (1988). A meta-analysis of the effects of direct instruction in special education. *Education and Treatment of Children, 11,* 364–374.

Wilson, H. K., Pianta, R. C., & Stuhlman, M. (2007). Typical classroom experiences in first grade: Climate and functional risk in the development of social competencies. *Exceptional Children, 108,* 81–96.

Wu, H. (2001). How to prepare students for algebra. *American Educator, 25*(2), 10–17.

Wu, H. (2008). Fractions, decimals, and rational numbers. Retrieved from University of California, Berkeley, Department of Mathematics, Hung-Hsi Wu's website: http://math.berkeley.edu/~wu/

Wu, H. (2009, February). *From arithmetic to algebra.* Paper presented at the Mathematicians Workshop Series Schedule, University of Oregon, Eugene.