Examining the Association Between Explicit Mathematics Instruction and Student Mathematics Achievement

Authors

Christian T. Doabler, Ph.D., Scott K. Baker, Ph.D., Derek B. Kosty, B.S., Keith Smolkowski, Ph.D., Ben Clarke, Ph.D., Saralyn J. Miller, Ph.D., Hank Fien, Ph.D.

Full Reference

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal, 115*, 303–333. doi: 10.1086/679969

Funding Source

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants **R305K040081** and **R305A080699** to the Center on Teaching and Learning at the University of Oregon

Abstract

Explicit instruction is a systematic instructional approach that facilitates frequent and meaningful instructional interactions between teachers and students around critical academic content. This study examined the relationship between student mathematics outcomes and the rate and quality of explicit instructional interactions that take place during core mathematics instruction in kindergarten classrooms using a multifaceted observation system. A total of 379 observations were conducted in 129 classrooms, involving approximately 2,200 students, across a 2-year span. Results suggest that the rate and quality of instructional interactions is related to student mathematics achievement. Implications for instruction and observation research are discussed.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305K040081 and R305A080699 to the Center on Teaching and Learning at the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Examining the Association Between Explicit Mathematics Instruction and Student Mathematics Achievement

The research community's current focus on developing and testing interventions that improve student outcomes has lead to a renewed interest and support for research that tries to specify the relations between instructional variables and student achievement. That is, there is interest not only in identifying programs, policies, and practices that increase student outcomes, but also in specifying the underlying mechanisms that are associated with those outcomes. For example, instructional interactions between teachers and students are a defining characteristic of classroom instruction and a component carefully defined in many education interventions (Cohen, Raudenbush, & Ball, 2003; Pianta & Hamre, 2009). The purpose of this paper is to examine the relation between the explicit instructional interactions that occur between teachers and students during kindergarten core mathematics instruction and student mathematics achievement. To measure this relationship, we employed a multifaceted, direct observation system that consisted of a low-inference instrument and two moderate-inference instruments. Based on this focus, the next section reviews two relevant research literatures: explicit mathematics instruction and observational research on effective instruction.

Explicit Mathematics Instruction And Its Role In Mathematics Proficiency

Explicit instruction is a systematic instructional approach in which ambiguity regarding the roles of teachers and students is minimized (Archer & Hughes, 2010; Hudson & Miller, 2006; Stein, Silbert, & Carnine, 2006). As Carnine, Silbert, Kame'enui, and Tarver (2004) observed, it is an instructional methodology for teaching foundational concepts, principles, and skills in the most "effective and efficient manner possible" (p. 5). Explicit instruction is most well known for its role in small-group interventions. However, applications of it in core instruction settings are also common (Baker, Fien, & Baker, 2010). When orchestrated well, explicit instruction is an effective approach for helping improve students' opportunities for long-term academic success.

For example, there is consensus among researchers about the value of explicit instruction for students with or at-risk for mathematics disabilities (MD). Mathematics intervention studies

consistently demonstrate that students with or at-risk for MD demonstrate greater gains in classrooms that provide explicit instruction compared to other instructional approaches (Baker, Gersten, & Lee, 2002; Gersten et al., 2009; Kroesbergen & Van Luit, 2003; National Mathematics Advisory Panel [NMAP], 2008; Swanson & Hoskyn, 1998; White, 1988). Two recent meta-analyses have summarized much of this work. One targeted mathematics interventions for students with MD (Gersten et al., 2009), and the second targeted students who struggled with mathematics but were not identified with MD (Baker et al., 2002).

Gersten et al. (2009) analyzed 41 studies targeting students with MD. Interventions were coded on seven dimensions including, (a) explicit instructional techniques, (b) the use of visual representations of quantitative relations, (c) student verbalization of mathematics concepts and strategies for solving problems, (d) attention to the range and sequence of examples used during instruction, (e) frequent assessment feedback to teachers and students, and (f) peer assisted instruction. In the explicit instruction studies teachers demonstrated step-by-step routines for solving problems and then students applied these routines to solve similar problems. Of the seven dimensions, explicit instruction had the largest impact, and at $g = 1.22$, 95% CI [0.78, 1.67], the magnitude of the effect was substantial.

When mathematics interventions are used with students at risk for mathematics difficulties but without identified disabilities, the value of explicit instruction is also clear, although the impact may not be as substantial. Baker et al. (2002) analyzed 15 intervention studies with students at risk for mathematics difficulties but without identified disabilities. These studies were coded according to 5 intervention categories: (a) providing data to teachers or students about mathematics performance, (b) peer tutoring/peer assisted mathematics instruction, (c) using parents to support classroom instruction, (d) explicit instruction, and (e) computer-assisted instruction. The effect for explicit instruction was medium to large, $d = 0.65$, 95% CI [0.40, 0.77], second in magnitude to providing feedback to students, $d = 0.71$, CI [0.27, 0.87].

The value of explicit instruction for students who are *not* at risk for mathematics difficulties is not as clear as it is for students who are struggling. The NMAP (2008) identified eight

methodologically rigorous studies that did not focus specifically on at-risk students. The primary contrast in these studies was the benefit of teacher-directed approaches (more explicit instruction) versus student-centered approaches (much less explicit instruction). Results were mixed, and the panel stated that the current research base does not lead to "any conclusive result about the value of student-centered instructional strategies in comparison to teacher-directed instructional strategies" (p. 6-24). It is worth noting that in all but one of these studies the control condition was teacher-directed or explicit instruction. Given that the control condition in most intervention studies typically receives less guidance and implementation support than the treatment condition (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005), it may be that the impact of explicit instruction is somewhat underestimated in these studies.

Explicit Instructional Interactions During Early Mathematics Instruction

A focal point of explicit instruction is frequent and purposefully planned instructional interactions among teachers and students around critical academic content. The provision of frequent instructional interactions, though necessary, is not sufficient by itself to facilitate mathematical proficiency. Instructional interactions must also be of high-quality. High-quality instructional interactions are those appropriately and sufficiently distributed across students and the stages of learning (e.g., acquisition, independent practice; Pianta & Hamre, 2009).

Characterizing high-quality, explicit instructional interactions are three key components: (a) clear and concise teacher demonstrations, (b) frequent opportunities for students to practice what teachers demonstrate, and (c) timely academic feedback from teachers to students related to students' attempts to solve academic problems. Teacher demonstrations are defined as a teacher's explanations, clarifications, and overt demonstrations involved in completing a step or series of steps in an academic problem. The purpose of such demonstrations is for teachers to clearly show students what they want them to do during a particular activity or task (Archer & Hughes, 2010; Baker et al., 2010). In mathematics, teachers can model lower level skills, such as solving number combinations, as well as higher order content, such as thinking aloud for students and showing them how to complete all of the steps necessary to solve mathematical

word problems. For example, a teacher might provide a vivid, step-by-step demonstration of the counting-up and the counting-down strategies to solve subtraction problems.

Student practice is another key component of explicit instructional interactions. Evidence from a variety of research lines suggests that the frequency of practice has important implications for improving outcomes in academic domains and performance-based disciplines (Ericsson et al., 2007; Swanson & O'Connor, 2009). Findings from early reading research studies, for example, document the association between the rate of practice opportunities and student reading achievement (Smolkowski & Gunn, 2012; Vadasay, Sanders, & Peyton, 2005). Practice helps students acquire new knowledge, retain previously learned material, build fluency or automaticity, and connect existing background knowledge with new and more sophisticated content (Pellegrino & Goldman, 1987; Prawat, 1989). A student response is synonymous with student practice. When teachers elicit student responses they facilitate opportunities for students to engage in learning and mastering academic content (Archer & Hughes, 2010; Brophy & Good, 1986; Simmons et al., 2011; Sutherland, Alder, & Gunter, 2003). An important category of student responses in mathematics, particularly in the early grades, involves mathematical discourse or math verbalizations. Accumulating evidence documents the relation between verbalizing math content and learning (Gersten et al., 2009; Sutherland et al., 2003). Mathematical verbalizations permit students to interact with the teacher and peers around critical mathematics content. Specifically, verbalizing can be viewed as a way to process and practice math content and in this manner becomes a critical component for supporting early development of mathematical proficiency.

During explicit mathematics instruction, teachers prompt groups of students as well as individuals to communicate and demonstrate their mathematical knowledge. Group responses serve as a mechanism for maintaining student engagement during an entire lesson. When group responses occur in unison, they provide a way for teachers to get a quick assessment of how well all students are grasping specific content (Blackwell & McLaughlin, 2005; Carnine et al., 2004). For example, a teacher might provide an opportunity for the entire class to count out 12 objects.

Similarly, a group of students might simultaneously state how the commutative property applies when adding two numbers.

Individual responses during mathematics instruction are defined as one student demonstrating his or her mathematical knowledge. These responses can be particularly effective when interspersed with questions that teachers target to the group at large. Individual responses allow students to solve problems and answer questions on their own and give teachers a clear way to determine whether specific individual students understand important content (Archer & Hughes, 2009; Carnine et al., 2009). When individual response opportunities are judiciously distributed across the classroom, teachers are able to monitor student progress and differentiate instruction for struggling learners.

Accompanying student practice within an explicit instructional interaction framework is academic feedback. Academic feedback is a teacher behavior defined as response affirmations and error corrections (Archer & Hughes, 2010; Doabler et al., 2012). Teachers provide academic feedback to extend learning opportunities and reduce potential misconceptions. In the context of explicit instructional routines, academic feedback is intended to be immediate and directly aligned with the preceding student response. For example, if a group of students misidentified a geometric shape, the teacher would first correct the mistake by stating the shape's name. Then, to complete the feedback cycle, the teacher would immediately ask the group to re-identify the shape. With that, students receive corrective feedback and one or more additional opportunities to practice solving the problem successfully.

Documenting Explicit Instructional Interactions through Direct Observation

In having teachers demonstrate what students are to do and having students practice those behaviors in the presence of the teacher, and getting feedback on their efforts, it is clear that a key characteristic of explicit instructional interactions are their public, observable nature. This characteristic lends interpretability through direct observation (Shavelson et al., 1986; Snyder et al., 2006). Researchers have begun to develop and validate a variety of direct observation

systems designed to estimate the value instructional interactions have on student outcomes (Brophy & Good, 1986; Connor, Morrison, & Petrella, 2004; Englert, 1984; Greenwood, Carta, Kamps, & Delquadri, 1995; Pianta & Hamre, 2009; Smolkowski & Gunn, 2012; Vaughn & Briggs, 2003). Many studies use moderate to high inference instruments, which rely on observer impressions to rate the quality of such interactions (Englert, Tarrant, & Mariage, 1992; Pianta & Hamre, 2009).

For example, Gersten and colleagues (Baker, Gersten, Haager, & Dingle, 2006; Gersten, Baker, Haager, & Graves, 2005; Haager, Gersten, Baker, & Graves, 2003) developed a moderate-inference observation instrument that targeted the quality of interactions of early reading instruction for English learners (EL). This instrument (Baker et al., 2006) requires observers with considerable expertise in early reading and ELs to assign ratings to 29 items after observing an entire reading lesson. Items are grouped into 7 clusters including explicit instruction. In one study involving first grade ELs in high poverty classrooms (Baker et al., 2006), the correlation between explicit instruction and reading proficiency was approximately .70.

Pianta and colleagues have developed another type of moderate to high-inference instrument called the *Classroom Assessment Scoring System* (CLASS; Howes et al., 2008; La Paro et al., 2009; Mashburn et al., 2008; Pianta & Hamre, 2009; Pianta, La Paro, & Hamre, 2008) that targets three domains: Emotional Supports, Classroom Organization, and Instructional Supports. Across these domains, observers rate 10 dimensions on a 7-point rating scale. Coding occurs approximately every 30 minutes, and observations last a minimum of two hours and sometimes span an entire school day. The CLASS instrument has been used in more than 4,000 classrooms nationwide and considerable scientific evidence indicates that the instrument demonstrates inter-rater reliability and provides a valid measure of instructional quality (Howes et al., 2008; La Paro et al., 2009; Mashburn et al., 2008; Pianta & Hamre, 2009). Evidence from two recent studies suggests a statistically significant relationship between the interactions captured by the CLASS and student math achievement (Howes et al., 2008; Mashburn et al., 2008).

Research studies have also relied on relatively low inference measurement approaches (Englert, 1984), where frequency and duration estimates are used to operationalize the interactions that occur between teachers and students. Relative to moderate and high inference instruments, low inference instruments capture these interactions at a more molecular level and are better poised to minimize observer inference. Low inference systems have served multiple roles in intervention studies. Primarily, researchers have used them to record the occurrences of observable behaviors that are mutually exclusive within a research study (e.g., Sutherland et al., 2003). Others have extended this purpose and advocated for their use in quantifying treatment intensity (Gersten, Baker, & Lloyd, 2000; Warren, Fey, & Yoder, 2007).

Recently, Smolkowski and Gunn (2012) examined the relationship between the rate of explicit instructional interactions and students' early reading achievement using a low-inference instrument. Their instrument, *Classroom Observation of Student-Teacher Interactions* (COSTI), estimates the link between observed instruction and student outcomes by measuring critical elements of explicit instruction that occur during overt instructional interactions between teachers and students around the foundational skills of early literacy. The COSTI targets general information about the learning environment and measures the frequency of four teacher-student behaviors: (a) teacher demonstrations, (b) student independent practice opportunities, (c) student errors, and (d) teacher-provided academic feedback. These behaviors are intended to represent a cyclical sequence of teacher-student interactions that occur in many explicit instruction routines. These routines are organized around an instructional sequence that begins with the teacher demonstrating a learning objective. Students then practice that objective through guided support from the teacher. As students gain initial proficiency with the objective, support is systematically withdrawn to increase opportunities for students to independently demonstrate what they have learned. When students make errors during independent practice, the teacher provides feedback to correct errors immediately and then resumes the instructional routine, making sure to provide additional practice on the types of items that prove difficult for students.

Smolkowski and Gunn (2012) conducted a validation study of the COSTI instrument in 54 kindergarten classrooms, involving a total of 235 observations across multiple years. The study focused on estimating the reliability of the COSTI, the stability of the observed behaviors across different observation time points, and the association between observed teacher and student behaviors and student reading achievement. Generalizability coefficients, represented by intraclass correlation coefficients (ICCs), indicated that observers reliably used the COSTI. ICCs ranged from .61 to .99, which represent substantial to nearly perfect inter-observer reliability (Landis & Koch, 1977). Findings also showed that most observed behaviors remained stable across observation occasions, with teacher demonstrations showing the greatest variability across time and student practice opportunities showing the greatest stability. In predicting student outcomes, three of the four observed behaviors (a) rates of student practice opportunities, (b) student errors, and (c) teacher–provided feedback (as well as the proportion of practice opportunities followed by an error) all predicted reading outcomes in the hypothesized directions (e.g., more practice would be associated with better reading outcomes; a smaller proportion of student errors would be associated with better reading outcomes). Overall, student practice opportunities demonstrated the strongest association with student reading achievement. Interestingly, rate of teacher demonstrations was not associated with student outcomes. According to Smolkowski and Gunn (2012), this lack of association may have been attributable to the fact that the COSTI measures the frequency of teacher demonstrations but not the quality of those demonstrations.

Purpose of the Study

The primary purpose of the current study was to examine the relation between the rate of explicit instructional interactions in kindergarten classrooms and student mathematics achievement using a low-inference observation instrument. Accumulating research in early reading documents the relation between frequent instructional interactions and improved student reading outcomes (Cooke, Galloway, Kretlow, & Helf, 2011; Nelson-Walker et al., 2013; Smolkowski & Gunn, 2012). Less is known in mathematics. In part this is because mathematics

has received less attention than reading in education research (Gersten, Clarke, & Mazzocco, 2007). In addition, of the research that has examined instructional interactions in early mathematics, the majority has been conducted in the context of small-group interventions (Gersten et al., 2009; Sutherland et al., 2003). To our knowledge, no large-scale research studies have systematically examined instructional interactions in the context of core mathematics instruction.

Core mathematics instruction is commonly thought of as the educational setting in which the majority of students, including those with or at risk for MD, receive instruction in the general education curriculum (Fuchs & Vaughn, 2012). In this whole classroom setting, the general education teacher typically uses a published core or basal mathematics program, such as *Math Expressions, Saxon Math,* or *Everyday Mathematics*, to deliver instruction on the mathematics standards addressed at each grade (Wu, 2011). The educational importance of core mathematics instruction is widely supported. Strong core instruction has, for example, implications for promoting mathematical proficiency and reducing long-term mathematics difficulties (Clarke et al., 2011; Fuchs & Vaughn, 2012). For most students, including those at risk for MD, it serves as the main source of mathematics instruction (Powell, Fuchs, & Fuchs, 2013). In kindergarten these implications are perhaps even more profound. Core mathematics instruction in kindergarten is typically children's first exposure to formal mathematics instruction and thus it represents a prime window of opportunity to put children on an early path for mathematics success. In the current study, we consider the value of explicit instructional interactions facilitated during kindergarten core mathematics instruction for improving the mathematics outcomes of all kindergarten students, including students with and without MD.

A secondary purpose of the current study was to examine the relation between the quality of explicit instructional interactions and student mathematics achievement using two moderate inference instruments. Moderate inference instruments permit an observer to concurrently collect information on both the quantitative and qualitative dimensions of instructional practice and student learning. Moreover, there is evidence that moderate and high inference instruments

provide information about effective instruction that may not be captured through low inference instruments. Several studies have shown that information documented by moderate inference instruments correlates higher with achievement than frequency measures of instructional activities (Gersten, Carnine, Zoref, & Cronin, 1986; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004; Stoolmiller, Eddy, & Reid, 2000).

We posed three research questions. First, what is the association between frequency-based components of explicit instructional interactions and student mathematics achievement? Second, what is the association between ratings of instruction quality and student mathematics achievement? Third, does the addition of the ratings of instruction quality predict gains in student math achievement over and above frequency-based components of explicit instructional interactions?

Method

Data for this study were obtained from a larger randomized control trial designed to investigate the efficacy of a kindergarten intervention curriculum, *Early Learning in Mathematics* (ELM; Clarke et al., 2011). The ELM efficacy trial included 129 kindergarten classrooms from 7 school districts and 46 schools in two different geographical regions: Oregon and Texas. Of the 129 classrooms, 68 and 61 were randomly assigned to treatment (i.e., ELM curriculum) and control conditions (i.e., standard district mathematics instruction), respectively. While the larger trial tested the efficacy of the ELM curriculum (for results of condition effects of ELM, see Clarke et al., 2011), the present study does not conduct tests of experimental condition. Rather, the present study focuses on examining the relationship between the rate and quality of explicit instructional interactions and student mathematics achievement. Thus, the present study includes classroom observation data and student outcomes from the 129 classrooms. The primary unit of analysis for the present study is the classroom.

Participants

Among the 46 schools, 32 were public schools, 11 were private, and 3 were charter schools. Student demographic data were available for only the 32 public schools. Policies of the private and charter schools prevented release of such data. In the 32 public schools, an average of 76% of the student population qualified for free or reduced lunch programs. The breakdown by student ethnicity in the Oregon schools was Hispanic (36%), Black (2%), White (56%), Asian and Pacific Islander (5%), American Indian (1%). In the Texas schools, the ethnicity breakdown was Hispanic (69%), Black (29%), White (1%), Asian and Pacific Islander (<1%), American Indian (<1%). Of the 129 classrooms, 64 and 65 were from schools in Oregon and Texas, respectively. A total of 112 provided a full-day kindergarten program and 17 provided a half-day program. All half-day classrooms were from schools in Oregon. One full-day classroom operated four days per week; otherwise the kindergarten programs ran five days per week. The sample consisted of 17 bilingual education classes; however all math instruction was conducted in English. Average class size was 20 students (SD = 3.7).

Classrooms were taught by 130 teachers, of which 129 held a teacher certification. One classroom had two teachers, each working a half-day schedule. All teachers participated for the duration of the study. Most teachers were female (98%) and had, on average, 5.5 years of teaching experience, and 4.1 years of experience teaching at the kindergarten level. Overall, 39% of the teachers held a graduate degree, and approximately 51% completed college-level coursework in Algebra. In this sample, 69% identified themselves as White, 20% as Hispanic, and 11% as representing another ethnic group.

Approximately 2,681 students began the school year in these 129 kindergarten classrooms. Oregon classrooms included 1,413 students and Texas classrooms included 1,268 students. Across the two years, 237 students dropped from the study (8.8% attrition), primarily because of family mobility. Approximately 209 students moved into the participating classrooms after the study began. Of the 2,681 students, 133 received special education services (5.6%) and 746 had limited English proficiency (29.8%), and 53% of the student sample was male. The sample used

for analysis on a published standardized outcome measure included 2,103 students at pretest and 2,270 students at posttest. The sample used for analysis on a set of curriculum-based measures included 2,202 students at pretest and 2,271 students at posttest.

Kindergarten Mathematics Instruction in the ELM Efficacy Trial

We observed core mathematics instruction in the 129 classrooms that participated in the efficacy trial (Clarke et al., 2011). In that study, classrooms in the treatment condition implemented the ELM curriculum, and classrooms in the comparison condition implemented standard district mathematics instruction. ELM is a full-year kindergarten mathematics curriculum designed for use in whole classroom settings. It includes 120 core lessons that address topics identified in the Common Core State Standards (2010). All classrooms in the study committed to teaching mathematics at least 45 minutes per day.

Mathematics instruction in the control condition included the use of a number of different published curricula and teacher-developed materials. Commonly used curricula in the control classrooms were *Everyday Mathematics, Bridges in Mathematics,* and *Houghton Mifflin*. The instructional focus and format varied, with some teachers focusing more on whole number concepts, and others focusing on particular aspects of geometry and measurement. Instruction was delivered through a variety of different mediums, including learning centers, small group activities, and whole-class delivered instruction.

Student Measures

Students were assessed at pretest and posttest on measures of foundational aspects of number proficiency (Gersten et al., 2012). The assessment battery included a general outcome measure of students' procedural and conceptual knowledge of whole numbers, and a set of early mathematics curriculum-based measures that focused on discrete skills of number proficiency. Trained staff administered all student measures, with data collection meeting acceptable reliability criteria (i.e., implementation fidelity of .95 or higher).

Test of early mathematics ability-third edition. The Test of Early Mathematics Ability-Third Edition (TEMA-3; Pro-Ed, 2007) is a standardized, 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 test-retest reliabilities of .97 and .82 to .93, respectively. For concurrent validity with other math outcome measures, the TEMA-3 manual reports coefficients ranging from .54 to .91. Standard scores were used in the analyses.

Early numeracy-curriculum-based measurement measures (EN-CBM). EN-CBM (Clarke & Shinn, 2004) is a set of measures based on principles of curriculum-based measurement (Shinn, 1989). Each 1-minute fluency-based measure assesses an important aspect of early numeracy development including oral counting, number identification, quantity discrimination, and strategic counting with strings of numbers. The EN-CBM measures have been validated for use with kindergarten students (Chard et al., 2005). The Oral Counting measure requires students to orally rote count as high as possible and the discontinue rule applies after the first counting error. The Number Identification measure requires students to orally identify numbers between 0 and 10 when presented with a set of printed number symbols. Quantity Discrimination requires students to name which of two visually presented numbers between 0 and 10 is greater. The Missing Number measure requires students to name the missing number from a string of numbers (0-10). Students are given strings of three numbers with the first, middle, or last number of the string missing. A total EN-CBM score, computed as the sum across all subtests, was used in subsequent analyses. We computed concurrent validity coefficients as the correlation between EN-CBM total scores and the TEMA-3 scores at pretest (r = .87) and posttest (r = .81). We computed test-retest reliability coefficients as the correlations

between adjacent assessment occasions (i.e., six weeks separated each assessment occasion). The average test-retest reliability was .89.

Classroom Observation of Student-Teacher Interactions–Mathematics

The Classroom Observation of Student-Teacher Interactions-Mathematics (COSTI-M) represents a modified version of the direct observation measure of reading instruction used by Smolkowski and Gunn (2012). Part of our rationale for adapting a reading observation instrument for use in kindergarten during core mathematics instruction is based on the nature of the development of foundational skills in the early grades. The type of measure Smolkowski and Gunn used may be sensitive to effective reading instruction in the early grades but less sensitive in later grades, after students have developed foundation skills. For example, effective instruction in phonological awareness or the alphabetic principle may be associated with how much teachers interact with students to produce the sounds in words and read words. Effective instruction in complex comprehension skills, however, might not be as highly associated with the frequency of explicit instructional interactions between teachers and students. For instance, more of the variance in comprehension growth may be associated with how much independent reading students engage in.

If this hypothesis is correct regarding overt reading interactions between teachers and students contributing more to reading growth in the later grades, it is reasonable to predict that the same principle might be operating in mathematics. That is, the importance of the frequency of instructional interactions between teachers and students might be relatively robust in the development of foundational skills with number sense and whole number and operations, and less influential as instruction shifts to students using foundational skills to solve increasingly complex mathematics problems (e.g., algebra).

Our rationale for adapting a reading observation system for use during mathematics instruction is also based on the overt, public nature of learning and instructional interactions in the early grades. Before students have developed the independent reading and writing skills

necessary to read and analyze grade level texts, they depend heavily on teachers to structure and guide learning opportunities through classroom discourse (Justice, Meier, & Walpole, 2005; McGinty, Justice, Piasta, Kaderavek, & Fan, 2011). This dynamic is also clear in early mathematics, for example, in which a relatively high percentage of the learning opportunities students encounter occurs through the medium of mathematical discourse. As demonstrated in previous research on explicit instructional interactions (Englert, 1984; Nelson-Walker et al., 2013; Smolkowski & Gunn, 2012; Sutherland et al., 2003), these discourse opportunities are typically managed by the teacher and optimally measured by direct observation.

The COSTI-M measures the frequency of explicit instructional interactions during kindergarten mathematics instruction. Specifically, it targets six instructional interaction behaviors, with two at the teacher level and four at the student level. All behaviors are coded in a continual, serial fashion and thus each behavioral occurrence is recorded. For teacher behaviors, observers coded teacher demonstrations and academic feedback. Teacher demonstrations included explanations, verbalizations of thought processes, and physical demonstrations. For example, observers coded a model if a teacher stated a math definition or demonstrated how to complete a multistep mathematical procedure. Academic feedback was a teacher's verbal reply or physical demonstration to a student response. Academic feedback took the form of either an error correction or a response affirmation, which was a change from the procedure used by Smolkowski and Gunn (2012), who only recorded corrective feedback.

At the student level, the current study differed from Smolkowski and Gunn (2012) by separating student practice opportunities at the individual and group levels and capturing both guided and independent response opportunities. In all, observers coded four student behaviors: (a) group responses, (b) individual responses, (c) errors, and (d) other forms of response. Group response was defined as a math related verbalization produced by two or more students. For example, a group response would consist of 15 students concurrently stating the name of a geometric shape. A group response would also be two students counting from 1 to 20. An

individual response was an opportunity for one student to verbalize or physically demonstrate her mathematical understanding and thinking. An individual response was coded when a teacher identified a specific student and asked her a math question (e.g., "Lucy, what shape is this?"). Observers also coded an individual response when the teacher posed a question to the group at large, in which it was implied an individual student would be asked to provide an answer or response (e.g., "Who can point to the numeral 5?" students raise hands and the teacher calls on one student to respond). To avoid coding extraneous conversation or answers that were not elicited by the teacher, the observation protocol required that both student group and individual responses be preceded by teacher-posed questions or requests.

Observers also coded students' verbal and physical response errors (e.g., counting incorrectly, pointing to an incorrect shape). Finally, observers used a code called "other" forms of responses to capture specific physical actions completed at the group level (i.e., by two or more students) in which the observer had difficulty confirming the accuracy of the response. Other forms of responses included group written exercises, use of math representations by multiple students, and peer-partner learning. In these instances, observers would code "other" responses due to the difficulty of judging response correctness. Specific examples are two students in very low voices counting by 5s to one another or a group of 18 students holding up three fingers to represent the numeral three. In these examples, observers would code one "other" response each.

In the current study, rates of COSTI-M behaviors were computed as the frequency of the following behaviors divided by the duration of observation in minutes: teacher demonstrations, group responses, individual responses, and other forms of student responses. Two conditional probabilities were also computed: the proportion of student responses in which students provided incorrect answers and the proportion of student responses followed by academic feedback. The mean rates of instructional interactions and the mean conditional probabilities across three observations per classroom were used as predictors in subsequent analyses.

Moderate Inference Instruments of Observed Instruction Quality

To account for the quality of teacher demonstrations and additional features of explicit instruction noted by Smolkowski and Gunn (2012), we designed two moderate inference instruments. Although both instruments were designed to complement the COSTI-M during the classroom observations, each had a different focus on instruction quality. One instrument measured of the quality of explicit instructional interactions narrowly, whereas the second instrument covered a much broader scope.

Quality of classroom instruction. The first instrument, the Quality of Classroom Instruction (QCI), was designed to measure the quality of explicit instructional interactions. Adapted from an instrument developed by Gersten and colleagues (Baker et al., 2006; Gersten, et al., 2005), we designed the QCI to complement the frequency-based COSTI-M and serve as a molecular measure of the quality of eight critical aspects of explicit instructional interactions. These aspects included teacher modeling, instructional pacing, response time, transitions between activities, student engagement, learning success, checks of student understanding, and academic feedback. Each aspect was rated on a 3-point rating scale. Observers completed the QCI at the conclusion of each observation conducted in the Oregon classrooms, relying on events that took place during instruction for scoring each item. Total QCI scores were computed as the mean across all items. The internal consistency of the QCI was high, with a coefficient alpha of .94. On the 3-point scale, a rating of 1 represented the lowest score and a rating of 3 represented the highest score. The mean across the three observations per classroom were used as instructional quality predictors in subsequent analyses

Ratings of classroom management and instructional support. The second instrument, the Ratings of Classroom Management and Instructional Support (RCMIS; Doabler & Nelson-Walker, 2009), was developed to serve as a broad measure of instruction quality and was used in tandem with the COSTI-M in Texas classrooms. The RCMIS was developed based on a critical analysis of several programs of observation research, including the work of Pianta and colleagues (Pianta & Hamre, 2009), Danielson (1996), and Englert and colleagues (Englert et al.,

1992). The RCMIS is comprised of 11 items that target general features of mathematics instruction quality, including classroom management techniques, delivery of instruction, and the learning environment. Internal consistency of the measure was high, .92 (coefficient alpha). To rate the quality of each item, observers used a 4-point rating scale, with scores of 1-2 representing the lower quality range and 3-4 representing the upper quality range. Observers relied on a detailed scoring rubric to differentiate between scores. Observers in the Texas classrooms completed the RCMIS at the conclusion of each classroom observation. Total RCMIS scores were computed as the mean across all items. The mean across the three observations per classroom were used as instructional quality predictors in subsequent analyses.

Observation Procedures

Classroom observations. Trained observers conducted observations in the fall (Round 1), winter (Round 2), and spring (Round 3) of each respective school year, with approximately six weeks separating each observation round. One observation per classroom in each round was planned and eight were missed due to scheduling conflicts or teacher absences. We considered the three observations at different time points as a snapshot for how kindergarten mathematics instruction may function across a given school year. Across the two years, a total of 379 observations were conducted; 7 of 129 were not completed in Round 1, 1 of 129 were not completed in Round 2, and 0 of 129 were not completed in Round 3. Of the total number of observations scheduled, 98% were conducted.

All classroom observations were scheduled in advance and conducted during the core mathematics instruction time period. Scheduled observations were not specific to mathematical content (e.g., whole numbers or geometry), lessons, or a particular instructional day (e.g., start or end of a weekly math unit). Observers remained in each classroom for the duration of mathematics instruction, with observations lasting between 30 and 90 minutes. Observations were conducted using the COSTI-M and one of two measures of instruction quality, depending on the region. Oregon observers used the QCI to rate instruction quality and observers in Texas

used the RCMIS. Observers used the COSTI-M during the entire instructional period and completed the RCMIS or QCI at the conclusion of each observation. In Oregon, the COSTI-M was used in all three of the observation rounds. In Texas classrooms, however, the COSTI-M was used in the winter and spring rounds only.

Observation training. Eighteen observers from Oregon and Texas conducted all classroom observations. The observers included former educators, doctoral students, faculty members, and experienced data collectors from a nonprofit research institute and a university located in the southwest. Observers received approximately 14 hours of training, with an initial training lasting six hours and two, 4-hour follow up trainings prior to the winter and spring observation rounds to help minimize observer drift and increase interobserver reliability. Training focused on direct observation procedures, kindergarten mathematics, and use of the observation instruments. Prior to observing classrooms on their own, observers were required to complete two reliability checkouts and meet an interobserver agreement criterion of .85 or higher on each checkout. The first was a video checkout, which had observers code a 5-minute video of kindergarten math instruction. Second, observers completed a real-time classroom checkout with a primary observer from the research team. All observers met the minimum interobserver agreement level for both checkouts.

Interobserver Reliability and Stability Estimates

COSTI-M. We measured interobserver reliability with intraclass correlation coefficients (ICC). On 74 occasions, two observers collected data simultaneously to assess interobserver reliability. The ICC gives the proportion of variance associated with the occasion, opposed to observers. We found ICCs of .67 for teacher models, .92 for group responses, .95 for individual responses, .91 for other forms of responses, .84 for errors, and .90 for feedback, all considered substantial to nearly perfect interobserver reliability (Landis & Koch, 1977).

Analogous to test-retest reliability, we estimated the stability of COSTI-M behaviors across time by calculating an ICC from three observations nested within each of the 129 classrooms.

The stability ICC represents the proportion of between-classroom variance out of the total variance comprised from the between- and within-classroom variance. The within-classroom variance provides an estimate of the day-to-day variability in instructional activities plus any unreliability in the measure, while the between-classroom variance captures those features of instruction that remain stable across time. Thus, a large proportion of within-classroom variance, indicated by a low ICC, suggests unstable behaviors, which require more observations to capture reliable estimates of their frequency across the school year. Conversely, higher ICCs suggest fewer observations are necessary to obtain a reasonable estimate of behavior across the school year. Rates of individual response opportunities and academic feedback were modestly stable over time, with ICCs of .34 and .35, respectively, suggesting that three observations per classroom may not provide adequate estimates of the true rates of these behaviors (Shoukri, Asyali, & Donner, 2004). Other COSTI-M behaviors were less stable: ICCs ranged from .13 to .19.

While the ICCs help explain how teacher behavior differs from day to day, they do not depend on the number of observations and do not describe the reliability of the observed mean, which represents the variables used in the analysis. The reliability of the observed mean depends on both the stability of behaviors and the number of observations per year. The reliability of observed mean across the three observations per teacher (Raudenbush & Bryk, 2002) were .60 and .62 for individual responses and academic feedback, respectively, which suggests that three observations adequately capture the mean rates of these behaviors. Reliabilities for the observed means of other COSTI-M behaviors were considerably lower, ranging from .30 to .41.

QCI and RCMIS. Interobserver reliability and estimates of stability for QCI and RCMIS were obtained by following the same procedures as described for the COSTI-M. Moderate to high interobserver reliability was obtained for QCI and RCMIS, with ICCs of .72 and .61, respectively. The measures were modestly stable over time, with stability ICCs of .35 and .33, respectively. The reliability of the observed mean QCI and RCMIS scores across the three

observations were .62 and .60, respectively.

Statistical Analyses

A series of multilevel models, with students nested within classrooms, tested whether the COSTI-M, QCI, and RCMIS predicted math achievement in kindergarten (Raudenbush & Bryk, 2002). We estimated these associations by regressing spring student-level math achievement scores (i.e., TEMA-3 and EN-CBM), adjusted for fall scores, on each predictor, separately. Analyses involving the COSTI-M utilized the combined sample (Oregon and Texas). QCI and RCMIS were collected in only Oregon or Texas, respectively, so tests of these predictors used their corresponding subsamples.

Multilevel modeling was conducted using HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004), and parameters were estimated using restricted maximum likelihood. To ease the interpretation of results, we computed Pseudo-R^2 (Singer & Willett, 2003) as a measure of effect size for each fixed effect of observer ratings on math achievement. Pseudo-R^2 represents the decrease in classroom-level variance between unconditional and conditional models, or the proportion of classroom-level variance explained in the outcome measure by a predictor or set of predictors. All *p*-values are two-tailed.

Results

Table 1 provides means, standard deviations, and sample sizes for each student outcome and observation measure by region and for the combined sample. Table 2 summarizes the intercorrelations between the observation-based measures used in this study. Although this study was conducted within the context of a larger efficacy trial (Clarke et al., 2011), experimental condition did not moderate the associations between the predictor variables and student outcomes. In addition, adding English Language learner and special education status as covariates did not change the results and, therefore, we did not include these variables in the statistical analyses.

< Tables 1 and 2>

Quantity Measures Of Instructional Interactions

We hypothesized that the rate of instructional interactions would predict student achievement in the spring of kindergarten, controlling for fall scores. Tables 3 and 4 summarize the multilevel models that tested whether instructional interactions predicted spring TEMA-3 and EN-CBM scores, adjusted for fall scores. As seen in Tables 3 and 4, respectively, rates of individual responses statistically significantly predicted covariate-adjusted spring TEMA-3 scores (p = .004, Pseudo-R^2 = .08) and EN-CBM scores (p = .017, Pseudo-R^2 = .05). Higher rates of individual responses were associated with better covariate-adjusted spring outcomes.[1]

In terms of teacher instructional behaviors, rates of demonstrations and the proportion of responses followed by academic feedback did not produce statistically significant results. Regarding student behaviors, rates of group responses, other forms of responses, and the proportion of responses in which students provided incorrect answers did not produce statistically significant results.

< Tables 3 and 4 >

Ratings Of Instruction Quality

We also hypothesized that the QCI and RCMIS measures would predict student achievement in the spring of kindergarten, controlling for fall scores. Table 5 summarizes the multilevel models that tested whether Oregon's instructional quality measure (i.e., QCI) predicted spring TEMA-3 and EN-CBM scores, adjusted for fall scores. As can be seen in the columns labeled Model 1 of Table 5, the QCI measure statistically significantly predicted covariate-adjusted spring TEMA-3 scores (p = .014, Pseudo-R^2 = .14) but not EN-CBM scores (p = .199, Pseudo-R^2 = .03). Table 6 summarizes similar results for the Texas instructional quality measure (i.e., RCMIS). As can be seen in the columns labeled Model 1 of Table 6, RCMIS statistically

[1] Rates of individual responses were not statistically significantly associated with the total amount of instruction time observed (r = -.04, p = .626). We also tested whether the associations between individual response rate and student achievement were a function of the duration of observation. The individual response rate by duration interactions were not statistically significant with respect to the TEMA-3 or EN-CBM (p = .625 and .608, respectively).

significantly predicted covariate-adjusted spring TEMA-3 scores (p = .039, Pseudo-R^2 = .05) but not EN-CBM scores (p = .626, Pseudo-R^2 = -.01). Across both states, higher quality of instructional interactions was associated with better covariate-adjusted spring TEMA-3 scores.

< Tables 5 and 6 >

Unique Associations Of Instructional Quality

We hypothesized that the QCI and RCMIS measures would account for unique variance in the covariate-adjusted spring TEMA-3 scores over and above the statistically significant instructional interaction predictors captured by the COSTI-M (i.e., rate of individual responses). Prior to evaluating these associations, we determined that region of study (i.e., Oregon vs. Texas) did not moderate the relationship between individual response rate and covariate-adjusted spring TEMA-3 (p = .256) or EN-CBM scores (p = .808), allowing us to proceed with subsequent analyses within region without additional interpretive caution. Region-specific associations between individual response rate are reported in the columns labeled Model 2 of Tables 5 and 6 for comparison purposes, but the best estimates of these associations appear in Tables 3 and 4, wherein the complete sample and the most information available was utilized for analysis.

Unique associations between student outcomes and the QCI and RCMIS after accounting for individual response rate are summarized in Tables 5 and 6 within the columns labeled Model 3. As can be seen in Table 5, the unique association between QCI on covariate-adjusted spring TEMA-3 scores after accounting for individual response rate was statistically significant (p = .05, Pseudo-R^2 = .08). As can be seen in Table 6, after accounting for individual response rate, the RCMIS was no longer a statistically significant predictor of covariate-adjusted spring TEMA-3 scores. Also seen in Table 6, we found trend-level and statistically significant unique associations between individual response rate, and covariate-adjusted spring TEMA-3 (p = .071, Pseudo-R^2 = .06) and EN-CBM scores (p = .036, Pseudo-R^2 = .06) after accounting for the RCMIS ratings.

These findings demonstrate two things. First, the relationship between individual response rate and covariate-adjusted spring EN-CBM scores remains positive and statistically significant after controlling for instructional quality. In other words, individual response rate predicts math achievement over and above instructional quality (the converse of our hypothesis). Second, although the relationship between individual response rate and covariate-adjusted spring TEMA-3 scores controlling for instructional quality is not statistically significant, the trend is positive.

Discussion

A primary aim of this study was to test the relationship between the rate of explicit instructional interactions and student mathematics achievement. In an explicit instructional approach, these interactions entail the teacher demonstrating what students will learn and providing academic feedback to students when they engage in learning activities requested by the teacher. Explicit instructional interactions also provide students with opportunities to verbalize and demonstrate their mathematical thinking and understanding of critical concepts and skills. In this final section, we briefly review key findings of the study, describe limitations, and discuss implications for instruction and future observation research.

Quantity Measures Of Instructional Interactions

The results were mixed among the quantity measures targeting instructional interactions. One finding aligned with our prediction showed fairly strong evidence supporting the frequency of response opportunities for individual students to verbalize and physically demonstrate their mathematical knowledge and thinking during interactions with the teacher. Results suggest that this type of student response is associated with student achievement on proximal and distal measures of mathematics. We also predicted that the same type of association would occur for student–teacher interactions in which teachers posed questions for a group of students to respond to, in contrast to a response expectation of an individual student. However, the frequency of group responses to teacher questions was not associated with achievement. Also, we did not find

that the frequency of teacher models or demonstrations—that is, teachers showing students what they expected them to be able to do—was associated with achievement gains.

Although preliminary, the finding demonstrating the association between individual student responses and learning gains is both meaningful and encouraging. It is meaningful because it complements previous work on the role of student responses in learning academic content. Our results suggest that increasing the rate of individual response opportunities during core mathematics instruction can help support early development of mathematical proficiency. Prior research has shown that opportunities for students to practice is a defining feature in a range of disciplines, including music, chess, and sports (Ericsson et al., 2007), and is pivotal in a number of academic areas, including early literacy (Nelson-Walker et al., 2013; Smolkowski & Gunn, 2012), beginning and later mathematics (Fuchs et al., 2010; Kilpatrick, Swafford, & Findell, 2001; Strickland & Maccini, 2010), beginning writing (Graham & Perin, 2007), and in communication and language development (Justice et al., 2005; McGinty et al., 2012; Warren et al., 2007).

The finding is encouraging because the methodology used in the current study offers a feasible way to determine the influence individual student response opportunities in core mathematics instruction have on student outcomes. Our work with a frequency-based observation instrument is the first attempt to estimate the relation between the rate of student response opportunities, both group and individual, and student mathematics achievement in the context of core instruction. Although our findings are based on instruction delivered to the entire class, this type of measurement tactic may be useful in documenting meaningful instructional interactions that occur in small-group mathematics interventions. Moreover, as evidenced by Smolkowski and Gunn (2012) and Nelson-Walker et al. (2013), the COSTI has utility in documenting the importance of student–teacher interactions in content areas other than mathematics. Additional research with the frequency-based measures such as the COSTI and COSTI-M is needed to examine how instructional interactions function during foundational skill

instruction delivered in the upper elementary and middle grades, and whether these interactions are related to gains in important reading and mathematics outcomes.

In terms of the rates of group responses and teacher demonstrations, our findings suggest that both were unrelated to achievement. Several reasons may explain these findings. One possibility concerns the fact that teachers showed substantial variability in how often they had groups of students respond to math-related questions during instruction—that is, to actively participate in the lesson. Experts suggest that managing whole-class discourse is among the most difficult challenges teachers face when teaching mathematics (Kilpatrick et al., 2001). Knowing when to initiate whole-class discussions and how to manage them requires strong pedagogical and content-related knowledge (Hill, Rowan, & Ball, 2005). It may be that the rate of using this type of group response approach matters only after teachers reach a certain level of proficiency in how to use this type of approach successfully in the first place.

A second reason why group responses were not associated with student outcomes may be because our definition of group responses requires further refinement. Instead of simply documenting group responses, it may be important to capture these types of response opportunities for large groups of students by factoring in the amount of instructional guidance teachers provide (Chard & Jungjohann, 2006). For example, group responses might be divided into two distinct categories: guided responses and independent responses (Archer & Hughes, 2010). A guided-group response would entail a teacher providing active guidance throughout the response cycle by, for example, responding along with the students. The teacher composing a number using place value blocks with an entire class is an example of a high level of teacher guidance during an instructional task. Conversely, an independent group response would be accomplished without any guidance from the teacher. For example, a teacher might ask a group of students to count from 1 to 20 and let the students do the task alone without counting along with them. It may be that more opportunities for groups of students to solve math problems on their own without close guidance from the teacher is correlated with outcomes. Regardless,

separating group responses by the level of guidance the teacher provides may help account for variance in in the association between rate of group responses and student mathematics achievement.

The complexity of teacher demonstrations may help explain why the frequency of models was not associated with student mathematics achievement. Our definition of a teacher model is an overt demonstration or explanation of a mathematical concept or skill. This definition, however, fails to take into account the complexity involved in teacher models. Describing the attributes and relative position of a three-dimensional shape is arguably more complex than identifying the name of the shape. Future observation research should attempt to simultaneously capture both the frequency and depth of complexity of teacher demonstrations and explanations. Accurately measuring these aspects of teacher models, however, will require a more sophisticated observation instrument than the type of paper-pencil approach we employed in this study. Technology-based observational systems, such as mobile platforms, have the potential to record multiple types of classroom events including variations in how teachers explain and demonstrate for students the types of problems they would like them to solve.

With respect to the proportion of student responses that are incorrect, as well as other quantitative dimensions of student responses, our findings revealed that these frequency-based components of explicit instruction were not significant predictors of learning. For example, the proportion of student responses in which teachers provide academic feedback was not significantly related to student mathematics outcomes. This is surprising given the fact that academic feedback is a hallmark of explicit instruction. In the current study, academic feedback was defined as either verifying correct responses or correcting student mistakes. One explanation for this non-significant finding is that the response affirmation side of academic feedback and teacher demonstrations may have looked quite similar to our data collectors, especially given the demands of coding rapidly occurring instructional events in real time. Further refinement of ways to uniquely quantify academic feedback should be explored.

Ratings of Instruction Quality

Additional contributions of the manuscript include an investigation of instruction quality in kindergarten mathematics classrooms and whether instruction quality accounts for unique variance over and above the frequency-based predictors. Smolkowski and Gunn (2012) noted that the COSTI does not take into account the quality of teacher demonstrations and student response opportunities. An important objective of this study was to examine whether moderate inference instruments of instruction quality (i.e., the QCI and RCMIS) contributed to estimating the association between early mathematics instruction and student mathematics achievement.

The results were also mixed in terms of the rated quality of instructional interactions. In both Oregon and Texas classrooms, ratings of instruction quality were found to be significant predictors of the primary outcome measure, TEMA-3, but not the EN-CBM, the secondary outcome measure. In Oregon, classrooms that provided higher quality core math instruction (e.g., teacher models, response opportunities) demonstrated greater gains on the TEMA-3. Similar results were found using the RCMIS in the Texas kindergarten classrooms. In keeping with previous research (Howes et al., 2008; La Paro et al., 2009), these findings suggest that the value of instruction quality, as rated by trained observers, is related to student academic outcomes.

Interestingly, our findings revealed that the QCI measure predicted performance on the TEMA-3, even after controlling for rates of individual response opportunities (the only quantitative predictor that was statistically significant). These findings suggest that the COSTI-M and QCI may complement each other by capturing different, albeit substantive, aspects of explicit instructional interactions. Whereas the COSTI-M records the quantity of instructional behaviors, the QCI focused on observers' impressions of the quality of such explicit teaching behaviors.

Conversely, results showed that the RCMIS did not remain statistically significant when accounting for rates of individual response opportunities. One possible explanation is that the RCMIS is a more general measure of instruction quality than the QCI and does not complement

the COSTI-M as well. The implication for this is that the RCMIS may require additional items that are more aligned with explicit instructional interactions.

Limitations of the Study

This study has several limitations. The mixed results, in which some observation measures were related to student outcomes and others were not, may be a function of low statistical power. Only two observations in the Texas classrooms were conducted using the COSTI-M and thus these two time points may have been too few to detect associations between some of the COSTI-M behaviors (e.g., group response opportunities) and student outcomes. We elected not to use the COSTI-M in the first round of observations in Texas so that observers could become more familiar with using the RCMIS to rate instruction quality.

Additionally, the decision to limit observations to three rounds across each school year and schedule one observation in each classroom per round was based on resources. Conducting classroom observations is expensive and in many cases additional issues, such as the distance separating participating schools and school schedules that are not designed for maximizing observation schedules (and rightly so), increases the already high financial burden. Nonetheless it remains that generally speaking many observations are needed to produce reliable estimates of complex classroom phenomena, such as quality of instruction.

The way we measured instruction across the year may have been a limitation in this study and may have contributed to the mixed findings. The reliability of the estimates of the COSTI-M behaviors (e.g., teacher models, group responses) was surprisingly low across the academic year. It is likely that these estimates will improve with more observations. We conducted three observations in each classroom, and managed this so that they occurred in the beginning, middle, and end of year. More than three observations may be necessary to obtain reliable estimates of the types of teacher and student behaviors we are interested in. Future research should also examine how multiple observations within a short timeframe impacts the reliability of the estimates of teacher behaviors. For example, researchers may be able to measure behaviors more

reliably by documenting instructional interactions across consecutive days within a single school week. We base this prediction on the structure of commercially-available core curriculums, which are often organized in instructional units that are taught across one or two-week periods (Bryant et al., 2008). The consistency of mathematics content coverage and instructional activities within these instructional units is likely to reduce the day-to-day variability in instructional interactions.

Finally, research is needed to investigate how content both within and across the mathematical domains of the Common Core State Standards (2010) influences the stability of observed behaviors across time. It may be, for example, that teaching practices are more stable when instruction focuses on skills associated with the mathematical domain of Counting and Cardinality, such as rational counting, compared to skills associated with the mathematical domain of Measurement and Data, such as measuring the attributes of an object. The fact that the mathematics experiences in elementary school classrooms focus primarily on whole number concepts and operations compared to other mathematical topics support this assumption (Rowan, Harrison, Hayes, 2004). That is, teachers' frequent experiences in teaching whole number concepts may increase the stability of instructional interactions. Moreover, the stability of instructional interactions may increase when instruction focuses on particular topics within a mathematical domain. For example, within the Operations and Algebraic Thinking domain, activities that target number combinations may produce more stable estimates of student response opportunities relative to activities on solving word problems. In this study, we did not standardize observation days based on mathematical content. For an observation to take place, all that was required was for the focus of instruction to be on mathematics related content (e.g., patterning, addition and subtraction). Future observation research could explore how frequency-based components of explicit math instruction and ratings of instruction quality might vary in form and function by mathematical domain (e.g., number and operations vs. geometry, vs. measurement).

Implications for Instruction

Although the results of this study were mixed, we believe the findings lend preliminary support for the importance of measuring, as we did, students interacting with teachers and peers around key math content during core mathematics instruction. The value of frequent and high quality instructional interactions on student mathematics achievement may be substantial. Previous research makes plain that student response opportunities are an essential component of good instruction and student learning. However, much more evidence is needed to determine what defines high quality student responses during mathematics instruction. It is unclear, for example, what amount and type of response opportunities students should receive and how the opportunities might change over the course of the school year and as students advance in grade. A need also exists for future research to determine if there is differential impact of basic and complex response opportunities on student mathematics achievement. The COSTI-M coding of student responses is efficient but does not capture all that is important about mathematics instruction. For example, the measure does not differentiate between foundational and higher-order student responses. Under the COSTI-M coding structure, simple response opportunities that require one-word answers (e.g., “Lucas, what does six plus one equal?”) and higher-order response opportunities that require more detailed explanations and justifications (e.g., “Miles, can you explain how you solved that addition problem?”) are similarly coded as individual responses. Including a coding scheme that documents how these types of student responses differ may better distinguish associations between the frequency of student responses and important mathematics outcomes.

Evidence is also needed on how student responses influence mathematics achievement when they work in concert with other aspects of explicit instruction, such as teacher demonstrations and academic feedback. For example, an intensive teaching episode might start with a teacher demonstrating how to solve a particular type of problem, followed by student response opportunities, first in response to the teacher’s request directed at the whole group of students,

and then in response to the teacher's requests directed at students individually. Interspersed within the teacher request–student response cycle would be feedback from the teacher offering encouragement, information about the accuracy and quality of responses, and in some cases elaborating on student answers as a way of further demonstrating how students can provide competent answers to math questions. At this point, empirical support is needed on how these explicit instructional behaviors should be structured and sequenced within episodes of early mathematics instruction. Answers to these kinds of questions could have implications for professional development, and in developing stronger curricular programs and mathematics interventions.

Implications for Observation Research

Our findings align with previous research (e.g., Pianta & Hamre, 2009; Smolkowski & Gunn, 2012) and support the proposition that the timing is right for standardized observation protocols to improve our understanding of the connection between the quantity and quality of evidence-based teaching practices and student achievement (Pianta & Hamre, 2009). Pianta and Hamre recently called for a substantial increase in rigorous research on the development and use of classroom observation instruments that target academic instruction. They suggested that standardized observation protocols could improve our understanding of instructional and environmental factors that are related to student achievement, and that this information could be used to improve professional development experiences for teachers that are tied directly to student learning. Pianta and Hamre argue that the timing is right for this because advances in theory and measurement, as well as interventions to improve student outcomes, make it possible to develop metrics of effective instruction and effective teachers that would not rely on "(a) the proxies of degrees or experience that bear only indirectly or not at all on student outcomes, nor (b) the tautology that effective teachers are those who produce achievement gains" (p. 109).

A robust classroom observation system could also play a critical role in measuring the cognitive demands of student responses to teacher requests. For example, one could argue that

having students verbally justify a solution method places greater cognitive load on learners than having them provide simple one-word answers, such as answering number combinations or identifying numbers. Observation systems that emphasize frequency of responses could be adapted to also include data related to the cognitive demand of the requests. It would be necessary, however, to use technology-based systems for data collection if data on the quality of specific teacher requests are going to be coded at the same time frequency information is collected.

Observational systems could also help researchers investigate and operationalize overall treatment intensity. For example, Warren et al., (2007) suggest that researchers use a frequency-based instrument to measure key variables of treatment intensity. Under this framework, researchers would estimate treatment intensity by capturing rates of specific *teaching episodes* or learning moments considered essential to student outcomes. The framework proposed by Warren et al. is different from the way other researchers have conceptualized treatment intensity (e.g., Bryant et al., 2011; Faggella-Luby & Deshler, 2008; Mellard, McKnight, & Jordan, 2010). Intensity is often conceptualized as a function of group size and factors of instructional time, such as the amount of time spent in each session, the number of days taught per week, and the total number of weeks. Although these aspects of instruction have a clear connection to the definition of treatment intensity defined by Warren et al., we think it is useful to precisely measure whether variables that should be influenced by group size, such as the number of instructional interactions a specific student might encounter in the span of 15 minutes, are in fact associated. For example, valid measures of instructional interactions would allow researchers to detect variability between and within experimental conditions. Consider a study in which researchers test the efficacy of a mathematics program and compare its effects against a control condition. If results favor the treatment condition but classrooms in both conditions were of similar size and provided similar amounts of instructional time per day, week, and school year, the researchers would be limited in the kinds of inferences drawn. Researchers could report that

the program improved achievement but would unable to say why and under what conditions achievement improved. The framework proposed by Warren et al. could lead to useful information about the active mechanisms of the treatment program.

Conclusion

In summary, as schools attempt to improve and accelerate student mathematics learning in order to meet the heightened expectations of new content standards, such as the Common Core State Standards Initiative (2010), it will become imperative to obtain a clear understanding of what works best and for whom in core mathematics instruction. Although preliminary, the results of this study demonstrate the potential importance of frequent, high-quality instructional interactions during core mathematics instruction in kindergarten classrooms. Future observational studies are needed, however, to further unpack the black box of core mathematics instruction.

References

Archer, A. L., & Hughes, C. A., (2010). *Explicit instruction—Effective and efficient teaching.* New York, NY: Guilford.

Baker, S., Fien, H., & Baker, D. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of Tier 1 and Tier 2 instructional supports. *Focus on Exceptional Children, 42*(9), 1–20.

Baker, S. K., Gersten, R. M., & Lee, D.-S. (2002). A synthesis of empirical research on teaching mathematics to low- achieving students. *Elementary School Journal, 103*, 51–73. doi:10.1086/499715

Baker, S. K., Gersten, R. M., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal, 107*, 199–220. doi: 10.1086/510655

Blackwell, A., & McLaughlin, T.F. (2005). Using guided notes, choral responding, and response card to increase student performance. *The International Journal of Special Education, 20,* 1-5.

Bryant, B. R., Bryant, D. P., Kethley, C., Kim, S. A., Pool, C.,& You-Jin, S. (2008). Preventing mathematics difficulties in the primary grades: The critical features of instruction in textbooks as part of the equation. Learning Disability Quarterly, 31, 21–35.

Bryant, D. P., Bryant, B. R., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. (2011). Early Numeracy Intervention Program for First-Grade Students With Mathematics Difficulties. *Exceptional Children, 78*(1), 7-23.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *The third handbook of research on teaching* (pp. 328–375). New York, NY: McMillan.

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

Chard, D.J., Clarke, B., Baker, S. K., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*, 3–14.

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.

Clarke, B., Smolkowski, K., Baker, S. K., Hank, F., Doabler, C.T., & Chard, D. J. (2011). The impact of a comprehensive Tier I core kindergarten program on the achievement of students at risk in mathematics. *Elementary School Journal*, *111*, 561–584.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25,* 119–142. doi: 10.3102/01623737025002119

Common Core State Standards Initiative. (2010). Common core state standards for mathematics. from http://www.corestandards.org/assets/CCSSI_Math Standards.pdf

Connor, C.M., Morrison, F.J., & Petrella, J.N. (2004). Effective reading comprehension instruction: Examining child x instruction interactions. *Journal of Educational*

Psychology, 96, 682-698.
Cooke, N. L., Galloway, T.W., Kretlow, A.G., Helf, S. (2011). Impact of the script in a supplemental reading program on instructional opportunities for student practice of specified skills. *The Journal on Special Education, 45*(1), 28–42. doi: 10.1177/0022466910361955
Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision & Curriculum Development.
Doabler, C.T., & Nelson-Walker, N.J. (2009). Ratings of Classroom Management and Instructional Support. Unpublished observation instrument, Center on Teaching and Learning, College of Education, University of Oregon, Eugene, OR.
Doabler, C.T., Strand-Cary, M., Jungjohann, K., Clarke, B, Fien, H., Baker, S., Smolkowski, K., & Chard, D. (2012). Enhancing core mathematics instruction for students at risk for mathematics disabilities. *TEACHING Exceptional Children, 44,* 48–57.
Englert, C. (1984). Effective direct instruction practices in special education settings. *Remedial and Special Education, 5,* 38–47. doi: 10.1177/074193258400500208
Englert, C. S., Tarrant, K. I., & Mariage, T. V. (1992). Defining and redefining instructional practice in special education: Perspectives on good teaching. *Teacher Education and Special Education, 15*, 62–86. doi: 10.1177/088840649201500203
Ericsson, K. A., Roring, R. W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies, 18*(1), 3–56. doi: 10.1080/13598130701350593
Faggella-Luby, M. N., & Deshler, D. D. (2008). Reading comprehension in adolescents with LD: What we know; what we need to learn. *Learning Disabilities Research & Practice, 23,* 70-78.
Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation research: A synthesis of the literature. Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network. nirn.fmhi.usf.edu/resources/publications/Monograph/index.cfm
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. *Exceptional Children, 76,* 135–156.
Fuchs, L.S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45,* 195-203.
Gersten, R. M., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial and Special Education, 26*, 197–214. doi: 10.1177/07419325050260040201
Gersten, R. M., Baker, S. K., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *Journal of Special Education, 34*(1), 2–18. doi: 10.1177/002246690003400101
Gersten, R., Carnine, D., Zoref, L., & Cronin, D. (1986). A multifaceted study of change in seven inner-city schools. *The Elementary School Journal*, *86*, 257–276.
Gersten, R. M., Chard, D., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*, 1202–1242. doi: 10.3102/0034654309334431

Gersten, R., Clarke, B., & Mazzocco, M. (2007). Historical and contemporary Perspectives on mathematical learning disabilities. In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 7–29). Baltimore: Brooks.

Graham, S., & Perin, D. (2007). A Meta-Analysis of Writing Instruction for Adolescent Students. *Journal of Educational Psychology, 99,* 445–476. doi: 10.1037/0022-0663.99.3.445

Greenwood, C. R., Carta, J. J., Kamps, D., & Delquadri, J. (1995). Ecobehavioral assessments systems software (EBHASS) practitioner's manual (Version 3.0). Kansas City: University of Kansas, Juniper Garden Children's Project.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42,* 371–406.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23,* 27–50.

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

Justice, L. M., Meier, J., & Walpole, S. (2005). Learning new words from storybooks: An efficacy study with at-risk kindergartners. *Language, Speech, and Hearing Services in Schools, 36*(1), 17–32.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: Mathematics Learning Study Committee, National Research Council.

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs: A meta-analysis. *Remedial & Special Education, 24*, 97–114. doi: 10.1177/07419325030240020501

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

Landis, R.J., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Mellard, D., McKnight, M., & Jordan, J. (2010). RTI tier structures and instructional intensity. *Learning Disabilities Research & Practice, 25,* 217-225.

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

National Mathematics Advisory Panel. (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. Washington, DC: US Department of Education □□□: 10.3102/0013189X08329195

Nelson-Walker, N.J., Fien, H., Kosty, D. B., Smolkowski, K., Smith, J.L.M., & Baker, S.K. (2013). Evaluating the effects of a systematic intervention on first-grade teachers' explicit reading instruction. *Learning Disability Quarterly.*

Pellegrino, J. W., & Goldman, S. R. (1987). Information processing and elementary mathematics. *Journal of Learning Disabilities, 20*(1), 23–32, 57.

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

Prawat, R. S. (1989). Promoting access to knowledge, strategy, and disposition in students: A research synthesis. *Review of Educational Research, 59*(1), 1–41.

Powell, S.R., Fuchs, L.S., & Fuchs, D. (2013). Reaching the mountaintop: Addressing the common core standards in mathematics for students with mathematics difficulties. *Learning Disabilities Research & Practice, 28,* 38-48.

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

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.,Jr (2000). *HLM 5: Hierarchical linear and nonlinear modeling*. (Statistical software manual). Skokie, IL: Scientific Software International.

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal, 105*, 103-107.

Schatschneider, C., Fletcher, J.M., Francis, D.J., Carlson, C.D., & Foorman, B.R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology, 96,* 265–282. doi: 10.1037/0022-0663.96.2.265

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

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

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, *13*, 251–271. doi: 10.1191/0962280204sm365ra

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *44*, 48–57.

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

Stein, M., Kinder, D., Silbert, J., & Carnine, D. (2006). *Designing effective mathematics instruction: A direct instruction approach*. Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

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

Strickland, T.K., & Maccini, P. (2010). Strategies for teaching algebra to students with learning disabilities: Making research to practice connections. *Intervention in School and Clinic,*

46, 38-45. Doi: 10.1177/1053451210369519

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

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

Vaughn, S., & Briggs, K. L. (2003). *Reading in the classroom: Systems for the observation of teaching and learning*. Baltimore, MD.: Paul H. Brookes Pub. Co.

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews, 13*, 70–77. doi: 10.1002/mrdd.20139

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. doi: 10.1177/105345129803300401

Wu, H. (2011). Phoenix rising: Bringing the common core state mathematics standards to life. *American Educator, 3*, 3-13.